UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1996

GRADUATE STUDENT RESEARCH PROGRAM FINAL REPORTS

VOLUME 9

ROME LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES

5800 Uplander Way

Culver City, CA 90230-6608

Program Director, RDL
Gary Moore

Program Manager, AFOSR
Major Linda Steel-Goodwin

Program Manager, RDL
Scott Licoscos

Program Administrator, RDL
Johnetta Thompson

Program Administrator, RDL
Rebecca Kelly

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Bolling Air Force Base

Washington, D.C.

December 1996

REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-00-

0739

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
December, 1996

3. REP[

4. TITLE AND SUBTITLE
1996 Summer Research Program (SRP), Graduate Student Research Program (GSRP), Final Reports, Volume 9, Rome Laboratory

5. FUNDING NUMBERS
F49620-93-C-0063

6. AUTHOR(S)
Gary Moore

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Research & Development Laboratories (RDL)
5800 Uplander Way
Culver City, CA 90230-6608

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research (AFOSR)
801 N. Randolph St.
Arlington, VA 22203-1977

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The United States Air Force Summer Research Program (USAF-SRP) is designed to introduce university, college, and technical institute faculty members, graduate students, and high school students to Air Force research. This is accomplished by the faculty members (Summer Faculty Research Program, (SFRP)), graduate students (Graduate Student Research Program (GSRP)), and high school students (High School Apprenticeship Program (HSAP)) being selected on a nationally advertised competitive basis during the summer intersession period to perform research at Air Force Research Laboratory (AFRL) Technical Directorates, Air Force Air Logistics Centers (ALC), and other AF Laboratories. This volume consists of a program overview, program management statistics, and the final technical reports from the GSRP participants at the Rome Laboratory.

14. SUBJECT TERMS
Air Force Research, Air Force, Engineering, Laboratories, Reports, Summer, Universities, Faculty, Graduate Student, High School Student

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank).*

Block 2. Report Date. Full publication date including day, month, and year, if available
(e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C - Contract	**PR** - Project
G - Grant	**TA** - Task
PE - Program Element	**WU** - Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with....; Trans. of....; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE - See authorities.
NASA - See Handbook NHB 2200.2.
NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.
DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA - Leave blank.
NTIS - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only).*

Blocks 17. - 19. Security Classifications. Self- explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

PREFACE

Reports in this volume are numbered consecutively beginning with number 1. Each report is paginated with the report number followed by consecutive page numbers, e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3.

This document is one of a set of 16 volumes describing the 1996 AFOSR Summer Research Program. The following volumes comprise the set:

VOLUME	TITLE
1	Program Management Report
	Summer Faculty Research Program (SFRP) Reports
2A & 2B	Armstrong Laboratory
3A & 3B	Phillips Laboratory
4	Rome Laboratory
5A, 5B & 5C	Wright Laboratory
6	Arnold Engineering Development Center, Wilford Hall Medical Center and Air Logistics Centers
	Graduate Student Research Program (GSRP) Reports
7A & 7B	Armstrong Laboratory
8	Phillips Laboratory
9	Rome Laboratory
10A & 10B	Wright Laboratory
11	Arnold Engineering Development Center, United States Air Force Academy, Wilford Hall Medical Center, and Wright Patterson Medical Center
	High School Apprenticeship Program (HSAP) Reports
12A & 12B	Armstrong Laboratory
13	Phillips Laboratory
14	Rome Laboratory
15A&15B	Wright Laboratory
16	Arnold Engineering Development Center

GSRP FINAL REPORT TABLE OF CONTENTS **i-x**

APPENDICIES:

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

INTRODUCTION

The Summer Research Program (SRP), sponsored by the Air Force Office of Scientific Research (AFOSR), offers paid opportunities for university faculty, graduate students, and high school students to conduct research in U.S. Air Force research laboratories nationwide during the summer.

Introduced by AFOSR in 1978, this innovative program is based on the concept of teaming academic researchers with Air Force scientists in the same disciplines using laboratory facilities and equipment not often available at associates' institutions.

The Summer Faculty Research Program (SFRP) is open annually to approximately 150 faculty members with at least two years of teaching and/or research experience in accredited U.S. colleges, universities, or technical institutions. SFRP associates must be either U.S. citizens or permanent residents.

The Graduate Student Research Program (GSRP) is open annually to approximately 100 graduate students holding a bachelor's or a master's degree; GSRP associates must be U.S. citizens enrolled full time at an accredited institution.

The High School Apprentice Program (HSAP) annually selects about 125 high school students located within a twenty mile commuting distance of participating Air Force laboratories.

AFOSR also offers its research associates an opportunity, under the Summer Research Extension Program (SREP), to continue their AFOSR-sponsored research at their home institutions through the award of research grants. In 1994 the maximum amount of each grant was increased from $20,000 to $25,000, and the number of AFOSR-sponsored grants decreased from 75 to 60. A separate annual report is compiled on the SREP.

The numbers of projected summer research participants in each of the three categories and SREP "grants" are usually increased through direct sponsorship by participating laboratories.

AFOSR's SRP has well served its objectives of building critical links between Air Force research laboratories and the academic community, opening avenues of communications and forging new research relationships between Air Force and academic technical experts in areas of national interest, and strengthening the nation's efforts to sustain careers in science and engineering. The success of the SRP can be gauged from its growth from inception (see Table 1) and from the favorable responses the 1996 participants expressed in end-of-tour SRP evaluations (Appendix B).

AFOSR contracts for administration of the SRP by civilian contractors. The contract was first awarded to Research & Development Laboratories (RDL) in September 1990. After

completion of the 1990 contract, RDL (in 1993) won the recompetition for the basic year and four 1-year options.

2. PARTICIPATION IN THE SUMMER RESEARCH PROGRAM

The SRP began with faculty associates in 1979; graduate students were added in 1982 and high school students in 1986. The following table shows the number of associates in the program each year.

YEAR	SRP Participation, by Year			TOTAL
	SFRP	GSRP	HSAP	
1979	70			70
1980	87			87
1981	87			87
1982	91	17		108
1983	101	53		154
1984	152	84		236
1985	154	92		246
1986	158	100	42	300
1987	159	101	73	333
1988	153	107	101	361
1989	168	102	103	373
1990	165	121	132	418
1991	170	142	132	444
1992	185	121	159	464
1993	187	117	136	440
1994	192	117	133	442
1995	190	115	137	442
1996	188	109	138	435

Beginning in 1993, due to budget cuts, some of the laboratories weren't able to afford to fund as many associates as in previous years. Since then, the number of funded positions has remained fairly constant at a slightly lower level.

3. RECRUITING AND SELECTION

The SRP is conducted on a nationally advertised and competitive-selection basis. The advertising for faculty and graduate students consisted primarily of the mailing of 8,000 52-page SRP brochures to chairpersons of departments relevant to AFOSR research and to administrators of grants in accredited universities, colleges, and technical institutions. Historically Black Colleges and Universities (HBCUs) and Minority Institutions (MIs) were included. Brochures also went to all participating USAF laboratories, the previous year's participants, and numerous individual requesters (over 1000 annually).

RDL placed advertisements in the following publications: *Black Issues in Higher Education*, *Winds of Change*, and *IEEE Spectrum*. Because no participants list either *Physics Today* or *Chemical & Engineering News* as being their source of learning about the program for the past several years, advertisements in these magazines were dropped, and the funds were used to cover increases in brochure printing costs.

High school applicants can participate only in laboratories located no more than 20 miles from their residence. Tailored brochures on the HSAP were sent to the head counselors of 180 high schools in the vicinity of participating laboratories, with instructions for publicizing the program in their schools. High school students selected to serve at Wright Laboratory's Armament Directorate (Eglin Air Force Base, Florida) serve eleven weeks as opposed to the eight weeks normally worked by high school students at all other participating laboratories.

Each SFRP or GSRP applicant is given a first, second, and third choice of laboratory. High school students who have more than one laboratory or directorate near their homes are also given first, second, and third choices.

Laboratories make their selections and prioritize their nominees. AFOSR then determines the number to be funded at each laboratory and approves laboratories' selections.

Subsequently, laboratories use their own funds to sponsor additional candidates. Some selectees do not accept the appointment, so alternate candidates are chosen. This multi-step selection procedure results in some candidates being notified of their acceptance after scheduled deadlines. The total applicants and participants for 1996 are shown in this table.

1996 Applicants and Participants			
PARTICIPANT CATEGORY	TOTAL APPLICANTS	SELECTEES	DECLINING SELECTEES
SFRP	572	188	39
(HBCU/MI)	(119)	(27)	(5)
GSRP	235	109	7
(HBCU/MI)	(18)	(7)	(1)
HSAP	474	138	8
TOTAL	1281	435	54

4. SITE VISITS

During June and July of 1996, representatives of both AFOSR/NI and RDL visited each participating laboratory to provide briefings, answer questions, and resolve problems for both laboratory personnel and participants. The objective was to ensure that the SRP would be as constructive as possible for all participants. Both SRP participants and RDL representatives found these visits beneficial. At many of the laboratories, this was the only opportunity for all participants to meet at one time to share their experiences and exchange ideas.

5. HISTORICALLY BLACK COLLEGES AND UNIVERSITIES AND MINORITY INSTITUTIONS (HBCU/MIs)

Before 1993, an RDL program representative visited from seven to ten different HBCU/Mis annually to promote interest in the SRP among the faculty and graduate students. These efforts were marginally effective, yielding a doubling of HBCI/MI applicants. In an effort to achieve AFOSR's goal of 10% of all applicants and selectees being HBCU/MI qualified, the RDL team decided to try other avenues of approach to increase the number of qualified applicants. Through the combined efforts of the AFOSR Program Office at Bolling AFB and RDL, two very active minority groups were found, HACU (Hispanic American Colleges and Universities) and AISES (American Indian Science and Engineering Society). RDL is in communication with representatives of each of these organizations on a monthly basis to keep up with the their activities and special events. Both organizations have widely-distributed magazines/quarterlies in which RDL placed ads.

Since 1994 the number of both SFRP and GSRP HBCU/MI applicants and participants has increased ten-fold, from about two dozen SFRP applicants and a half dozen selectees to over 100 applicants and two dozen selectees, and a half-dozen GSRP applicants and two or three selectees to 18 applicants and 7 or 8 selectees. Since 1993, the SFRP had a two-fold applicant

increase and a two-fold selectee increase. Since 1993, the GSRP had a three-fold applicant increase and a three to four-fold increase in selectees.

In addition to RDL's special recruiting efforts, AFOSR attempts each year to obtain additional funding or use leftover funding from cancellations the past year to fund HBCU/MI associates. This year, 5 HBCU/MI SFRPs declined after they were selected (and there was no one qualified to replace them with). The following table records HBCU/MI participation in this program.

SRP HBCU/MI Participation, By Year				
YEAR	SFRP		GSRP	
	Applicants	Participants	Applicants	Participants
1985	76	23	15	11
1986	70	18	20	10
1987	82	32	32	10
1988	53	17	23	14
1989	39	15	13	4
1990	43	14	17	3
1991	42	13	8	5
1992	70	13	9	5
1993	60	13	6	2
1994	90	16	11	6
1995	90	21	20	8
1996	119	27	18	7

6. SRP FUNDING SOURCES

Funding sources for the 1996 SRP were the AFOSR-provided slots for the basic contract and laboratory funds. Funding sources by category for the 1996 SRP selected participants are shown here.

1996 SRP FUNDING CATEGORY	SFRP	GSRP	HSAP
AFOSR Basic Allocation Funds	141	85	123
USAF Laboratory Funds	37	19	15
HBCU/MI By AFOSR (Using Procured Addn'l Funds)	10	5	0
TOTAL	188	109	138

SFRP - 150 were selected, but nine canceled too late to be replaced.
GSRP - 90 were selected, but five canceled too late to be replaced (10 allocations for the ALCs were withheld by AFOSR.)
HSAP - 125 were selected, but two canceled too late to be replaced.

7. COMPENSATION FOR PARTICIPANTS

Compensation for SRP participants, per five-day work week, is shown in this table.

1996 SRP Associate Compensation

PARTICIPANT CATEGORY	1991	1992	1993	1994	1995	1996
Faculty Members	$690	$718	$740	$740	$740	$770
Graduate Student (Master's Degree)	$425	$442	$455	$455	$455	$470
Graduate Student (Bachelor's Degree)	$365	$380	$391	$391	$391	$400
High School Student (First Year)	$200	$200	$200	$200	$200	$200
High School Student (Subsequent Years)	$240	$240	$240	$240	$240	$240

The program also offered associates whose homes were more than 50 miles from the laboratory an expense allowance (seven days per week) of $50/day for faculty and $40/day for graduate students. Transportation to the laboratory at the beginning of their tour and back to their home destinations at the end was also reimbursed for these participants. Of the combined SFRP and

GSRP associates, 65 % (194 out of 297) claimed travel reimbursements at an average round-trip cost of $780.

Faculty members were encouraged to visit their laboratories before their summer tour began. All costs of these orientation visits were reimbursed. Forty-five percent (85 out of 188) of faculty associates took orientation trips at an average cost of $444. By contrast, in 1993, 58 % of SFRP associates took orientation visits at an average cost of $685; that was the highest percentage of associates opting to take an orientation trip since RDL has administered the SRP, and the highest average cost of an orientation trip. These 1993 numbers are included to show the fluctuation which can occur in these numbers for planning purposes.

Program participants submitted biweekly vouchers countersigned by their laboratory research focal point, and RDL issued paychecks so as to arrive in associates' hands two weeks later.

In 1996, RDL implemented direct deposit as a payment option for SFRP and GSRP associates. There were some growing pains. Of the 128 associates who opted for direct deposit, 17 did not check to ensure that their financial institutions could support direct deposit (and they couldn't), and eight associates never did provide RDL with their banks' ABA number (direct deposit bank routing number), so only 103 associates actually participated in the direct deposit program. The remaining associates received their stipend and expense payments via checks sent in the US mail.

HSAP program participants were considered actual RDL employees, and their respective state and federal income tax and Social Security were withheld from their paychecks. By the nature of their independent research, SFRP and GSRP program participants were considered to be consultants or independent contractors. As such, SFRP and GSRP associates were responsible for their own income taxes, Social Security, and insurance.

8. CONTENTS OF THE 1996 REPORT

The complete set of reports for the 1996 SRP includes this program management report (Volume 1) augmented by fifteen volumes of final research reports by the 1996 associates, as indicated below:

1996 SRP Final Report Volume Assignments

LABORATORY	SFRP	GSRP	HSAP
Armstrong	2	7	12
Phillips	3	8	13
Rome	4	9	14
Wright	5A, 5B	10	15
AEDC, ALCs, WHMC	6	11	16

APPENDIX A – PROGRAM STATISTICAL SUMMARY

A. Colleges/Universities Represented

Selected SFRP associates represented 169 different colleges, universities, and institutions, GSRP associates represented 95 different colleges, universities, and institutions.

B. States Represented

SFRP -Applicants came from 47 states plus Washington D.C. and Puerto Rico. Selectees represent 44 states plus Puerto Rico.

GSRP - Applicants came from 44 states and Puerto Rico. Selectees represent 32 states.

HSAP - Applicants came from thirteen states. Selectees represent nine states.

Total Number of Participants	
SFRP	188
GSRP	109
HSAP	138
TOTAL	435

Degrees Represented			
	SFRP	GSRP	TOTAL
Doctoral	184	1	185
Master's	4	48	52
Bachelor's	0	60	60
TOTAL	188	109	297

SFRP Academic Titles	
Assistant Professor	79
Associate Professor	59
Professor	42
Instructor	3
Chairman	0
Visiting Professor	1
Visiting Assoc. Prof.	0
Research Associate	4
TOTAL	188

Source of Learning About the SRP		
Category	Applicants	Selectees
Applied/participated in prior years	28%	34%
Colleague familiar with SRP	19%	16%
Brochure mailed to institution	23%	17%
Contact with Air Force laboratory	17%	23%
IEEE Spectrum	2%	1%
BIIHE	1%	1%
Other source	10%	8%
TOTAL	100%	100%

APPENDIX B -- SRP EVALUATION RESPONSES

1. OVERVIEW

Evaluations were completed and returned to RDL by four groups at the completion of the SRP. The number of respondents in each group is shown below.

Table B-1. Total SRP Evaluations Received

Evaluation Group	Responses
SFRP & GSRPs	275
HSAPs	113
USAF Laboratory Focal Points	84
USAF Laboratory HSAP Mentors	6

All groups indicate unanimous enthusiasm for the SRP experience.

The summarized recommendations for program improvement from both associates and laboratory personnel are listed below:

A. Better preparation on the labs' part prior to associates' arrival (i.e., office space, computer assets, clearly defined scope of work).

B. Faculty Associates suggest higher stipends for SFRP associates.

C. Both HSAP Air Force laboratory mentors and associates would like the summer tour extended from the current 8 weeks to either 10 or 11 weeks; the groups state it takes 4-6 weeks just to get high school students up-to-speed on what's going on at laboratory. (Note: this same argument was used to raise the faculty and graduate student participation time a few years ago.)

2. 1996 USAF LABORATORY FOCAL POINT (LFP) EVALUATION RESPONSES

The summarized results listed below are from the 84 LFP evaluations received.

1. LFP evaluations received and associate preferences:

Table B-2. Air Force LFP Evaluation Responses (By Type)

		How Many Associates Would You Prefer To Get ?								(% Response)			
		SFRP				GSRP (w/Univ Professor)				GSRP (w/o Univ Professor)			
Lab	Evals Recv'd	0	1	2	3+	0	1	2	3+	0	1	2	3+
AEDC	0	-	-	-	-	-	-	-	-	-	-	-	-
WHMC	0	-	-	-	-	-	-	-	-	-	-	-	-
AL	7	28	28	28	14	54	14	28	0	86	0	14	0
FJSRL	1	0	100	0	0	100	0	0	0	0	100	0	0
PL	25	40	40	16	4	88	12	0	0	84	12	4	0
RL	5	60	40	0	0	80	10	0	0	100	0	0	0
WL	46	30	43	20	6	78	17	4	0	93	4	2	0
Total	84	32%	50%	13%	5%	80%	11%	6%	0%	73%	23%	4%	0%

LFP Evaluation Summary. The summarized responses, by laboratory, are listed on the following page. LFPs were asked to rate the following questions on a scale from 1 (below average) to 5 (above average).

2. LFPs involved in SRP associate application evaluation process:
   a. Time available for evaluation of applications:
   b. Adequacy of applications for selection process:
3. Value of orientation trips:
4. Length of research tour:
5 a. Benefits of associate's work to laboratory:
   b. Benefits of associate's work to Air Force:
6. a. Enhancement of research qualifications for LFP and staff:
   b. Enhancement of research qualifications for SFRP associate:
   c. Enhancement of research qualifications for GSRP associate:
7. a. Enhancement of knowledge for LFP and staff:
   b. Enhancement of knowledge for SFRP associate:
   c. Enhancement of knowledge for GSRP associate:
8. Value of Air Force and university links:
9. Potential for future collaboration:
10. a. Your working relationship with SFRP:
    b. Your working relationship with GSRP:
11. Expenditure of your time worthwhile:

(Continued on next page)

12. Quality of program literature for associate:
13. a. Quality of RDL's communications with you:
    b. Quality of RDL's communications with associates:
14. Overall assessment of SRP:

Table B-3. Laboratory Focal Point Reponses to above questions

	AEDC	*AL*	*FJSRL*	*PL*	*RL*	*WHMC*	*WL*
# Evals Recv'd	0	7	1	14	5	0	46
Question #							
2	-	86 %	0 %	88 %	80 %	-	85 %
2a	-	4.3	n/a	3.8	4.0	-	3.6
2b	-	4.0	n/a	3.9	4.5	-	4.1
3	-	4.5	n/a	4.3	4.3	-	3.7
4	-	4.1	4.0	4.1	4.2	-	3.9
5a	-	4.3	5.0	4.3	4.6	-	4.4
5b	-	4.5	n/a	4.2	4.6	-	4.3
6a	-	4.5	5.0	4.0	4.4	-	4.3
6b	-	4.3	n/a	4.1	5.0	-	4.4
6c	-	3.7	5.0	3.5	5.0	-	4.3
7a	-	4.7	5.0	4.0	4.4	-	4.3
7b	-	4.3	n/a	4.2	5.0	-	4.4
7c	-	4.0	5.0	3.9	5.0	-	4.3
8	-	4.6	4.0	4.5	4.6	-	4.3
9	-	4.9	5.0	4.4	4.8	-	4.2
10a	-	5.0	n/a	4.6	4.6	-	4.6
10b	-	4.7	5.0	3.9	5.0	-	4.4
11	-	4.6	5.0	4.4	4.8	-	4.4
12	-	4.0	4.0	4.0	4.2	-	3.8
13a	-	3.2	4.0	3.5	3.8	-	3.4
13b	-	3.4	4.0	3.6	4.5	-	3.6
14	-	4.4	5.0	4.4	4.8	-	4.4

3. 1996 SFRP & GSRP EVALUATION RESPONSES

The summarized results listed below are from the 257 SFRP/GSRP evaluations received.

Associates were asked to rate the following questions on a scale from 1 (below average) to 5 (above average) - by Air Force base results and over-all results of the 1996 evaluations are listed after the questions.

1. The match between the laboratories research and your field:
2. Your working relationship with your LFP:
3. Enhancement of your academic qualifications:
4. Enhancement of your research qualifications:
5. Lab readiness for you: LFP, task, plan:
6. Lab readiness for you: equipment, supplies, facilities:
7. Lab resources:
8. Lab research and administrative support:
9. Adequacy of brochure and associate handbook:
10. RDL communications with you:
11. Overall payment procedures:
12. Overall assessment of the SRP:
13. a. Would you apply again?
    b. Will you continue this or related research?
14. Was length of your tour satisfactory?
15. Percentage of associates who experienced difficulties in finding housing:
16. Where did you stay during your SRP tour?
    a. At Home:
    b. With Friend:
    c. On Local Economy:
    d. Base Quarters:
17. Value of orientation visit:
    a. Essential:
    b. Convenient:
    c. Not Worth Cost:
    d. Not Used:

SFRP and GSRP associate's responses are listed in tabular format on the following page.

Table B-4. 1996 SFRP & GSRP Associate Responses to SRP Evaluation

	Arnold	Brooks	Edwards	Eglin	Griffis	Hanscom	Kelly	Kirtland	Lackland	Robins	Tyndall	WPAFB	average
# res	6	48	6	14	31	19	3	32	1	2	10	85	257
1	4.8	4.4	4.6	4.7	4.4	4.9	4.6	4.6	5.0	5.0	4.0	4.7	4.6
2	5.0	4.6	4.1	4.9	4.7	4.7	5.0	4.7	5.0	5.0	4.6	4.8	4.7
3	4.5	4.4	4.0	4.6	4.3	4.2	4.3	4.4	5.0	5.0	4.5	4.3	4.4
4	4.3	4.5	3.8	4.6	4.4	4.4	4.3	4.6	5.0	4.0	4.4	4.5	4.5
5	4.5	4.3	3.3	4.8	4.4	4.5	4.3	4.2	5.0	5.0	3.9	4.4	4.4
6	4.3	4.3	3.7	4.7	4.4	4.5	4.0	3.8	5.0	5.0	3.8	4.2	4.2
7	4.5	4.4	4.2	4.8	4.5	4.3	4.3	4.1	5.0	5.0	4.3	4.3	4.4
8	4.5	4.6	3.0	4.9	4.4	4.3	4.3	4.5	5.0	5.0	4.7	4.5	4.5
9	4.7	4.5	4.7	4.5	4.3	4.5	4.7	4.3	5.0	5.0	4.1	4.5	4.5
10	4.2	4.4	4.7	4.4	4.1	4.1	4.0	4.2	5.0	4.5	3.6	4.4	4.3
11	3.8	4.1	4.5	4.0	3.9	4.1	4.0	4.0	3.0	4.0	3.7	4.0	4.0
12	5.7	4.7	4.3	4.9	4.5	4.9	4.7	4.6	5.0	4.5	4.6	4.5	4.6
Numbers below are percentages													
13a	83	90	83	93	87	75	100	81	100	100	100	86	87
13b	100	89	83	100	94	98	100	94	100	100	100	94	93
14	83	96	100	90	87	80	100	92	100	100	70	84	88
15	17	6	0	33	20	76	33	25	0	100	20	8	39
16a	-	26	17	9	38	23	33	4	-	-	-	30	
16b	100	33	-	40	-	8	-	-	-	-	36	2	
16c	-	41	83	40	62	69	67	96	100	100	64	68	
16d	-	-	-	-	-	-	-	-	-	-	-	0	
17a	-	33	100	17	50	14	67	39	-	50	40	31	35
17b	-	21	-	17	10	14	-	24	-	50	20	16	16
17c	-	-	-	-	10	7	-	-	-	-	-	2	3
17d	100	46	-	66	30	69	33	37	100	-	40	51	46

4. 1996 USAF LABORATORY HSAP MENTOR EVALUATION RESPONSES

Not enough evaluations received (5 total) from Mentors to do useful summary.

5. 1996 HSAP EVALUATION RESPONSES

The summarized results listed below are from the 113 HSAP evaluations received.

HSAP apprentices were asked to rate the following questions on a scale from 1 (below average) to 5 (above average)

1. Your influence on selection of topic/type of work.
2. Working relationship with mentor, other lab scientists.
3. Enhancement of your academic qualifications.
4. Technically challenging work.
5. Lab readiness for you: mentor, task, work plan, equipment.
6. Influence on your career.
7. Increased interest in math/science.
8. Lab research & administrative support.
9. Adequacy of RDL's Apprentice Handbook and administrative materials.
10. Responsiveness of RDL communications.
11. Overall payment procedures.
12. Overall assessment of SRP value to you.
13. Would you apply again next year? Yes (92 %)
14. Will you pursue future studies related to this research? Yes (68 %)
15. Was Tour length satisfactory? Yes (82 %)

	Arnold	Brooks	Edwards	Eglin	Griffiss	Hanscom	Kirtland	Tyndall	WPAFB	Totals
# resp	5	19	7	15	13	2	7	5	40	113
1	2.8	3.3	3.4	3.5	3.4	4.0	3.2	3.6	3.6	3.4
2	4.4	4.6	4.5	4.8	4.6	4.0	4.4	4.0	4.6	4.6
3	4.0	4.2	4.1	4.3	4.5	5.0	4.3	4.6	4.4	4.4
4	3.6	3.9	4.0	4.5	4.2	5.0	4.6	3.8	4.3	4.2
5	4.4	4.1	3.7	4.5	4.1	3.0	3.9	3.6	3.9	4.0
6	3.2	3.6	3.6	4.1	3.8	5.0	3.3	3.8	3.6	3.7
7	2.8	4.1	4.0	3.9	3.9	5.0	3.6	4.0	4.0	3.9
8	3.8	4.1	4.0	4.3	4.0	4.0	4.3	3.8	4.3	4.2
9	4.4	3.6	4.1	4.1	3.5	4.0	3.9	4.0	3.7	3.8
10	4.0	3.8	4.1	3.7	4.1	4.0	3.9	2.4	3.8	3.8
11	4.2	4.2	3.7	3.9	3.8	3.0	3.7	2.6	3.7	3.8
12	4.0	4.5	4.9	4.6	4.6	5.0	4.6	4.2	4.3	4.5
Numbers below are percentages										
13	60%	95%	100%	100%	85%	100%	100%	100%	90%	92%
14	20%	80%	71%	80%	54%	100%	71%	80%	65%	68%
15	100%	70%	71%	100%	100%	50%	86%	60%	80%	82%

DEVELOPMENT OF A USER-FRIENDLY COMPUTER ENVIRONMENT FOR BLIND SOURCE SEPARATION STUDIES

Parker E.C. Bradley
Department of Physics

Western Illinois University
1 University Circle
Macomb, IL 61455

Final Report for:
Graduate Student Research Program
Rome Labs

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Labs

September 1996

DEVELOPMENT OF A USER-FRIENDLY COMPUTER ENVIRONMENT FOR BLIND SOURCE SEPARATION STUDIES

Parker E.C. Bradley
Department of Physics
Western Illinois University

Abstract

This paper discusses the design of a digital array signal processing environment, geared toward investigations in blind source separation techniques. The system integrates the interactive numerical processing capabilities of Math Works' MATLAB[1] with multichannel data acquisition, and signal processing hardware. A user-friendly graphical user interface, utilizing MATLAB's graphical programming features, facilitates analysis of data, evaluation of algorithms, and modular addition of hardware/software components.

DEVELOPMENT OF A USER-FRIENDLY COMPUTER ENVIRONMENT FOR BLIND SOURCE SEPARATION STUDIES

Parker E.C. Bradley

Introduction

Sensor array technology promises an array (pun possibly intended) of applications from non-invasive data acquisition, signal recognition, and jammer resistance. A future key aspect of these areas, aside from spatial filtering and source localization, lies in the development of robust algorithms to perform blind source separation. The development and testing of such algorithms in a "real-life" scenario can be very difficult and time consuming, requiring specialized hardware and software tools that combine multichannel data acquisition/processing, and advanced numerical computation. As with any system that promotes effective algorithm testing and development, it is imperative that all the features be accessible via a user-friendly interface, which requires a minimum investment in training time.

An environment fulfilling this criteria is under development as depicted in Figure 1. The current exploratory setup consists of a host computer that provides overall control and analysis, a multichannel data acquisition unit, a signal conditioner, a sensor array (acoustic or electromagnetic), and assorted peripheral devices. The requirements of powerful numerical processing and a user-friendly environment have been satisfied by integrating array hardware with MATLAB. MATLAB offers advanced mathematical computation through built-in functions, [3] "plug-in toolboxes", and user defined functions [3][4]. It also has an easy to use interpreted programming language, is extensible with scripts and executable files, and has excellent graphical capabilities [1][2][3]. The rest of this paper describes the array hardware and software, and future directions of development to expand the usefulness of the system.

Hardware

The array has been set up to allow for both off-line algorithm development, and real-time signal processing of multichannel data. There are a number of subsystems comprising the hardware environment and include: a host computer, a sensor array (currently microphones, but could also be antennas for radio work), a signal conditioner, a signal collection/processing unit, and miscellaneous peripherals. Figure 1 provides a block diagram of the hardware components and their interconnections; the following paragraphs describe each in greater detail.

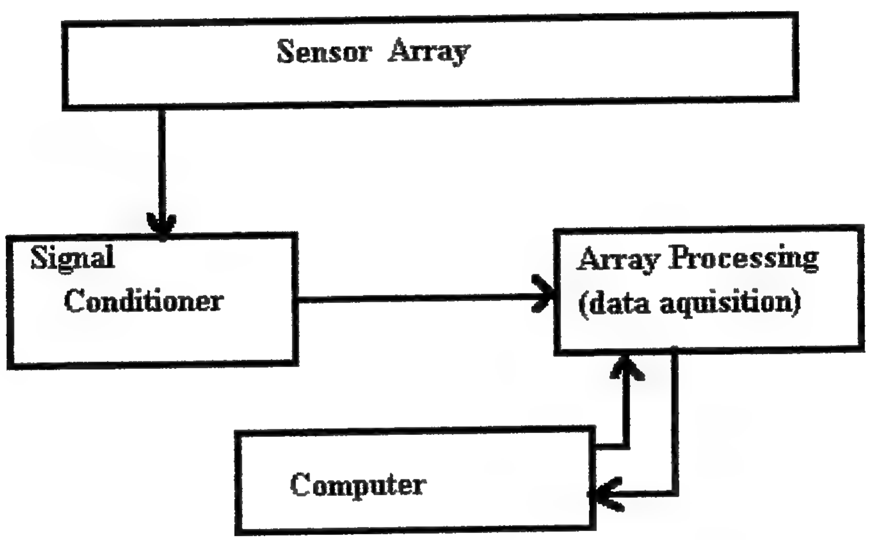


Figure 1: Current hardware and computer setup.

There are many benefits to using an "intelligent host" as the center of a research system: the ability to use the caveats of various operating systems, graphical displays, data input, mass storage, high-level languages, control and data sharing over high speed networks (or even more

remotely via phone lines, or packet radio links), and existing software (don't want to re-invent the wheel). For portability, and convenience, and since it was available, a Gateway Pentium notebook was used. The notebook ran Windows 95 [2] and was equipped with a variety of high quality software, including MATLAB. The computer's parallel port was utilized for interface to the array processing unit. As development (and funding) progresses an Alpha workstation and/or an Ultra Sparc may be interfaced to array processing hardware as well.

The array processing unit is the most specialized (and expensive) part of the system. In order to address portability concerns, an external analog to digital converter with digital I/O was chosen. The unit is manufactured by IOtech[3] and is called "WaveBook 512". The unit has eight differential inputs with a maximum sampling rate of one million samples per second. With eight channels in simultaneous use, a sampling rate of 125,000 samples per second is possible. The actual sampling rate achieved depends upon the number of channels utilized and the speed of the I/O port on the computer. To improve performance, a sample and hold unit can be added to the WaveBook 512. [5][1]

An array of eight unidirectional microphones is connected to an amplifier which in turn is connected to the array processing unit. To minimize noise and signal loss between devices, coaxial cable was used wherever possible. Signal sources can be comprised of speaker arrays, people talking, sounds from other animate or inanimate objects, pre-recorded files, or combinations thereof.

Further hardware development, see figure 2, begs for the use of an anechoic chamber with computer controlled stepper motors controlling sensor/emitter array geometries, and other

---

[1] WaveBook® is a trademark of IOtech®.

computer controlled stepper motors controlling sensor/emitter array geometries, and other controllable variables. Also it would be useful to have some degree of control, while not at the keyboard, therefore a remote control system would be of great benefit.

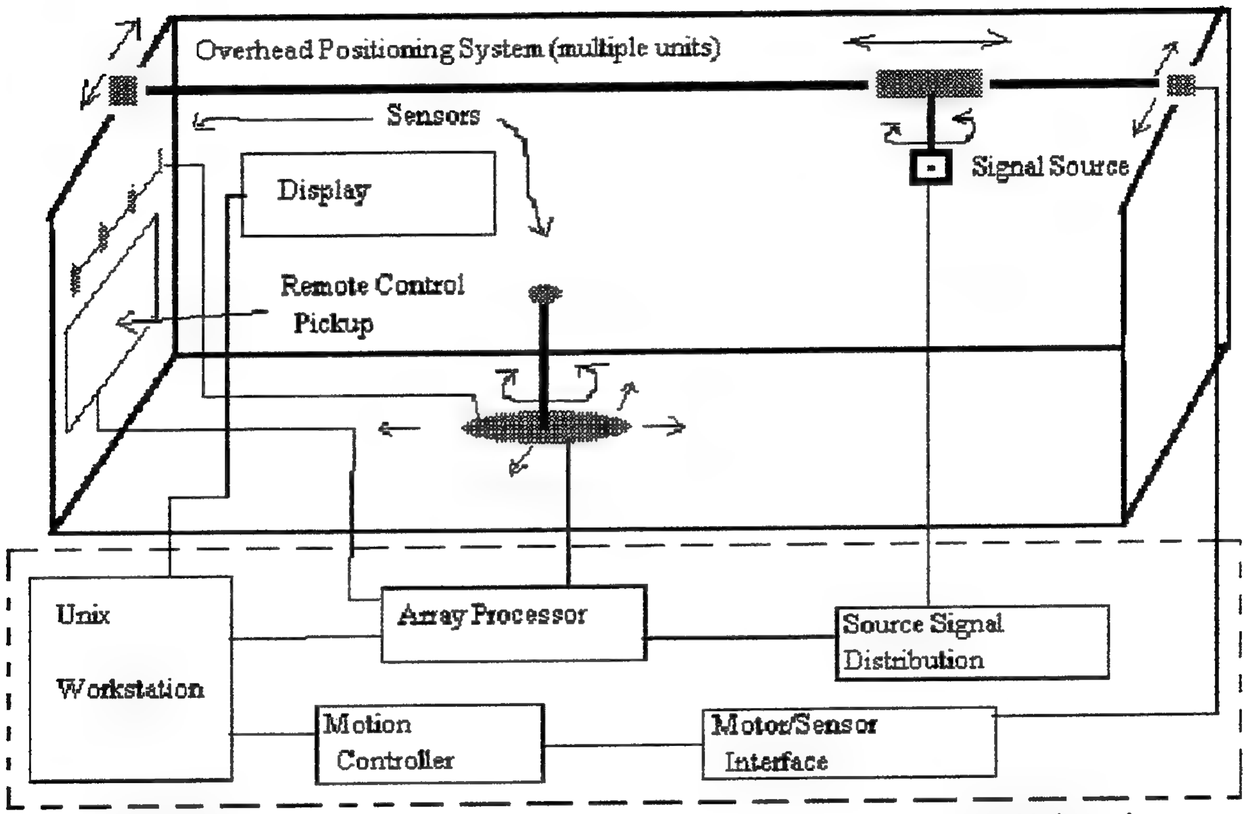


Figure 2:Suggested arrangement for possible further development of array research environment.

Software

The host computer (host is being used a little loosely here, and should not be applied in the strict sense of the word, as it would in a Unix system; although when a Unix workstation is finally integrated into the setup, the definition would follow to the letter) initializes the array processing unit and communicates via a server-client software model -- a "server" responds to requests by a "client", and the two exchange instructions and data following a simple protocol. Currently it is possible to vary data collection time, number of input channels, sampling rate, and initiate data collection/emission.

The host program, under development, provides an interface to the array processing unit and to various data analysis and record keeping tools, thus unifying the entire research and development process into a single application. The host program, affectionately referred to as BASIT: Blind Array Signal Identification Test-Bed, is a MATLAB script (M-file) which itself calls other MATLAB scripts. The host script interacts with the array processing unit through the use of MEX-files that dynamically link to C subroutines which have access to the full range of operating system I/O features. The MATLAB host script spawns the array interface and serves as a front end to the low-level MEX-files and provide the necessary command formatting and error checking [1][6].

Keeping in mind the goal of unifying the research/development process, a key aspect of the project is to develop a library of blind source separation algorithms which are known to be effective for given situations. These algorithms can then be called upon as needed, compared, contrasted, analyzed with a library of statistical/analytical tools, and possibly hybridized with other

methods.

Example

Figure 3 shows the results from a blind source separation algorithm, SOBI: Second Order Blind Identification, developed by Dr. Adel Belouchrani [7]. The results shown are for two signals "artificially" mixed by the computer. As you can see from figures 3-1,3-2, and 3-3 the algorithm works quite well in this situation.

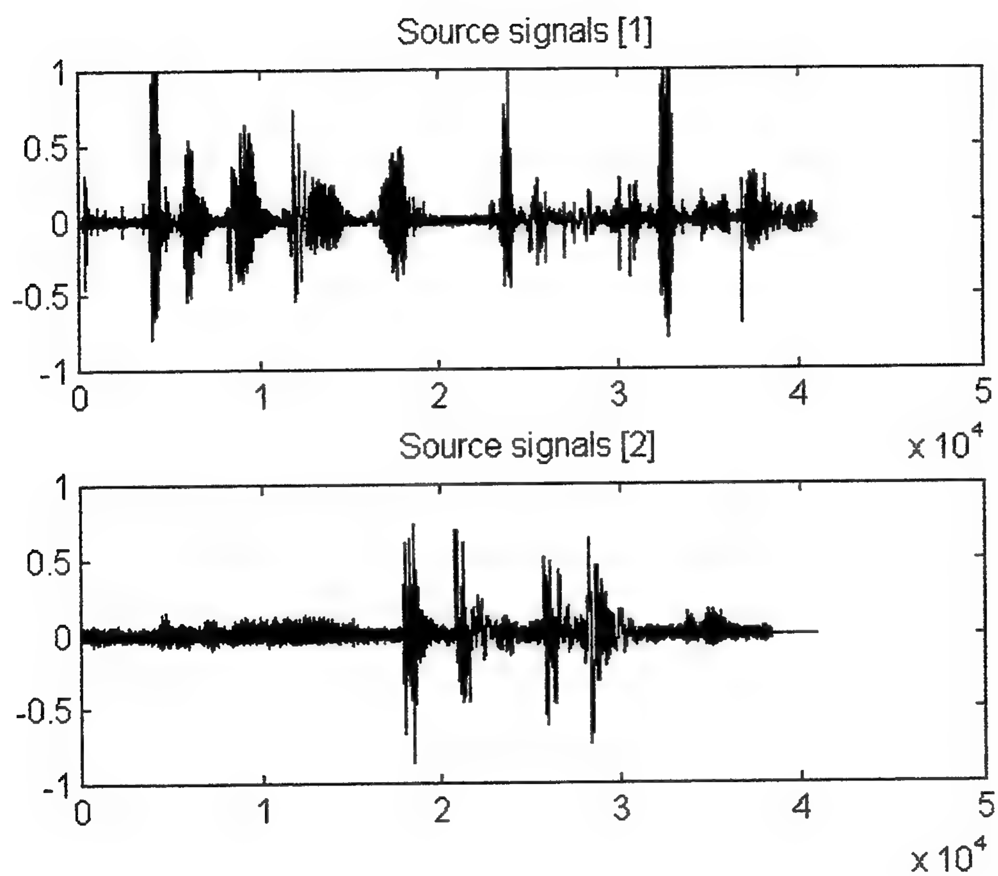


Figure 3-1: Source signals 1 & 2. The source signals are pre-recorded voice tracks sampled at 8,000 Hertz.

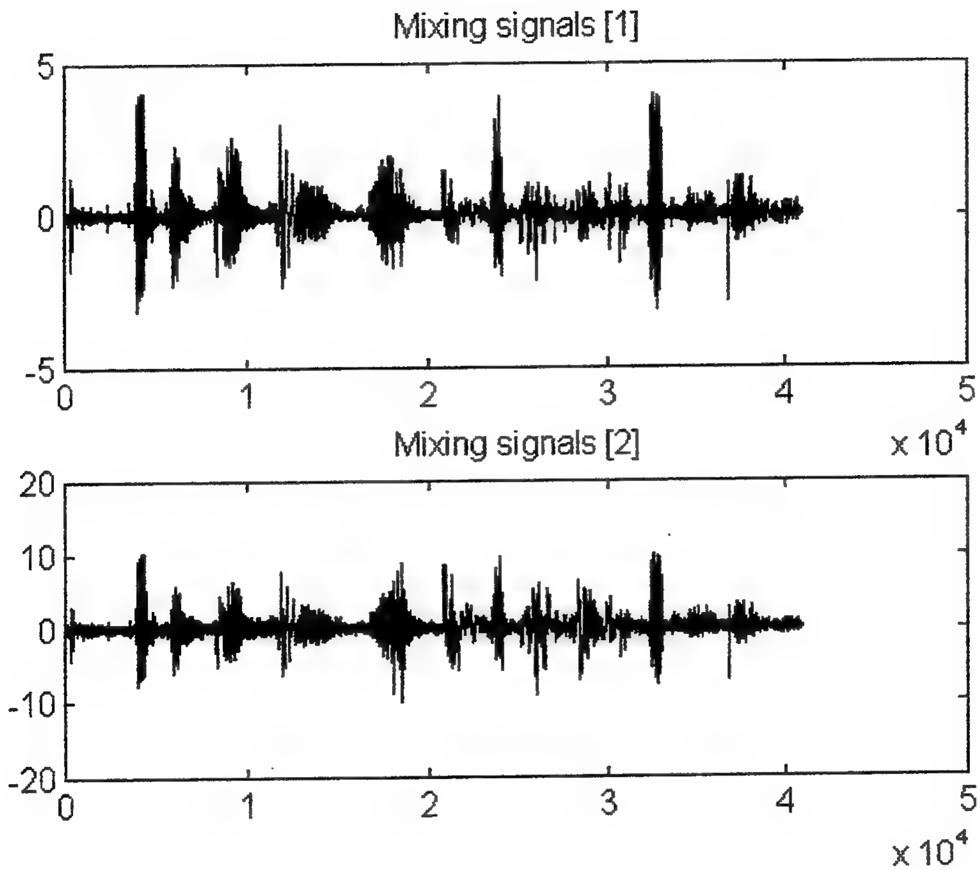


Figure 3-2: The respective mixing, by the computer, of the pre-recorded signals 1 & 2 as the mixture would appear to sensors 1 & 2 in a real situation, for a given sensor-array manifold.

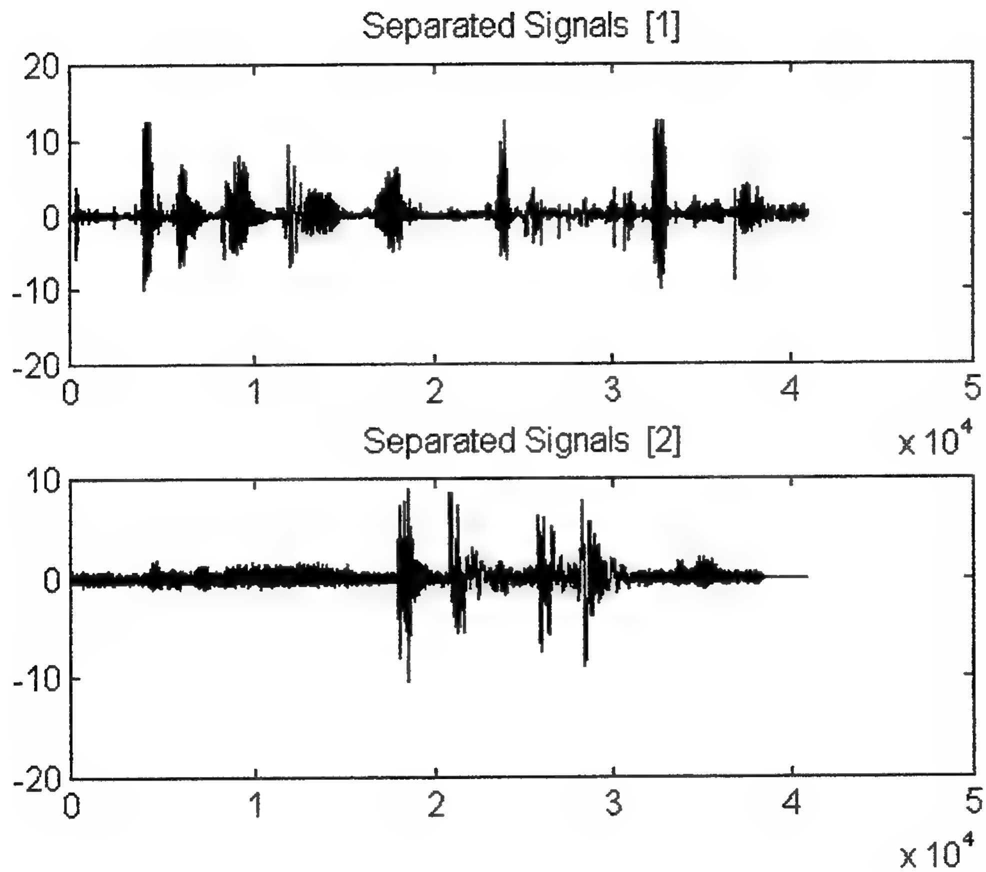


Figure 3-3: The signals that SOBI (a blind source separation algorithm) separated from the artificial mixture of signals. Notice the similarity to the original signals.

However, when the algorithm is presented with data mixed by the "real-world" it fails miserably . There were theoretical indications that this would be the case (due to the convolutive nature of sound, which the algorithm was not designed to handle), but as figures 4-1, 4-2, & 4-3 demonstrate, a real-world test drives the point home -- indicating other directions to follow such as convolutive and neural net approaches [8][9].

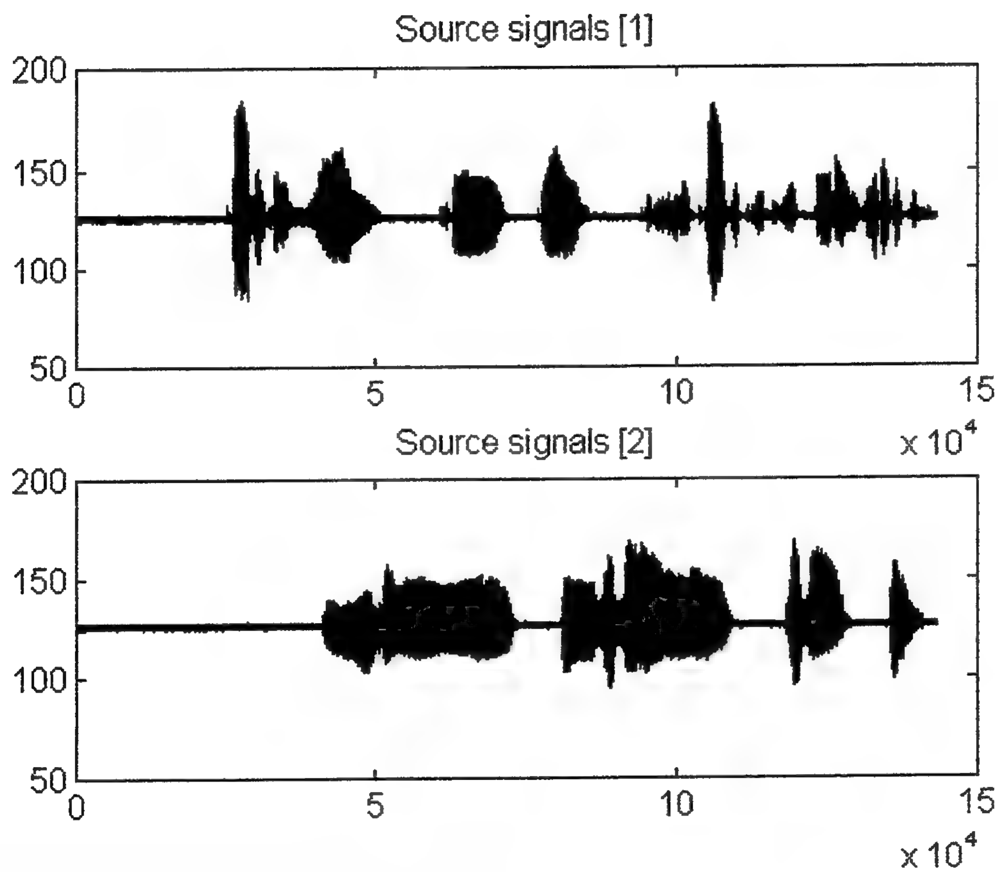


Figure 4-1: Real-world sound source signals.

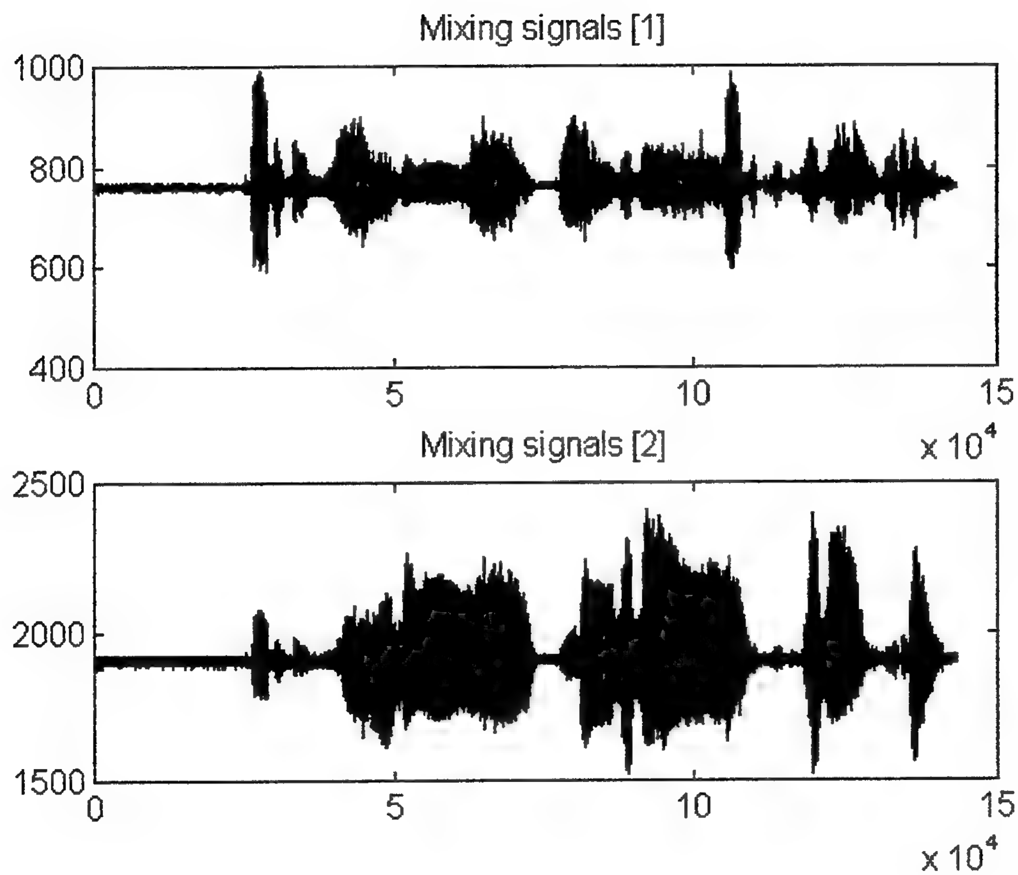


Figure 4-2: Real world mixing of two sound signals, from the perspectives of mics 1 & 2.

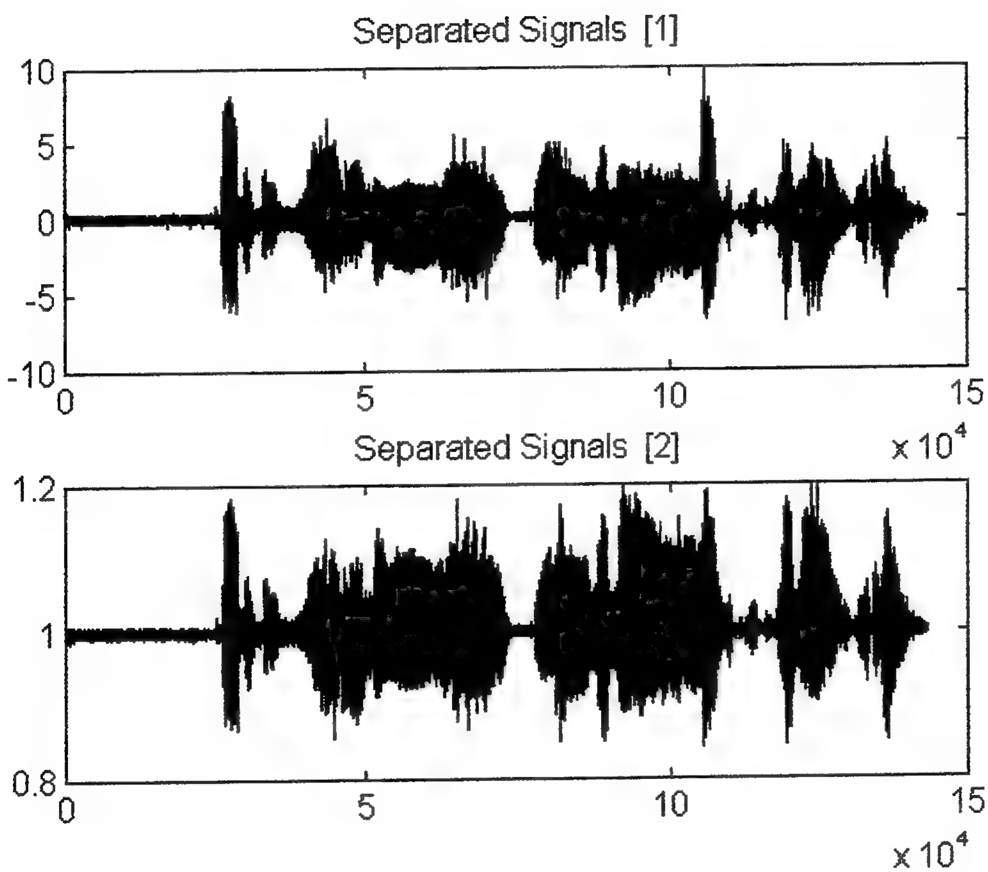


Figure 4-3: SOBI's (a blind source separation algorithm) attempt to separate signals from a real-world mixture. Compare to the originals, and notice that the attempt was unsuccessful.

Conclusion

An environment has been described that could provide a powerful, yet simple to use mechanism for experimental research on microphone and antenna arrays. While the creation of such a system is a large investment, the example presented, and common sense, indicate that this type of environment is essential for meaningful experimental research, not only in the area of blind source separation, but in array research as a whole.

References

[1] The Math Works, Inc., *Matlab External Interface Guide*, ©1996.

[2] The Math Works, Inc., *Building a Graphical User Interface,* © 1996.

[3] The Math Works, Inc., *Matlab Reference Guide,* © 1996.

[4] Hanselman, Duane, and Littlefield, Bruce, *Mastering Matlab,* Prentice Hall, © 1996.

[5] Iotech, *WaveBook 512 User's Manual,* © 1996.

[6] Borland, *Borland C++ Programmer's Guide*, © 1996.

[7] Belouchrani, A., Cardoso, J.F., Moulines, E., "Second Order Blind Separation of Temporally Correlated Sources", Paper received directly from author -- to be published.

[8] Thi Hoang-Lan Nguyen, and Jutten, Christian, "Blind Source Separation for Convolutive Mixtures", *Signal Processing*, vol. 45, 1995, pp. 209-229.

[9] Marsman Henkjan, "A Neural Net Approach to the Source Separation Problem -- Master's Thesis" *University of Twente,* Department of Electrical Engineering, Laboratory of Network Theory, P.O. Box 217, 7500 AE Enshede, The Netherlands. Report # EL-BSC 075N95.

A WEB BROWSER DATABASE INTERFACE
USING HTML AND CGI PROGRAMMING

Charles J. Harris
Graduate Student
Department of Computer Science

State University of New York Institute of Technology
P.O. Box 3050
Utica, New York 13504-3050

Final Report for:
Graduate Student Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

September 1996

A WEB BROWSER DATABASE INTERFACE USING HTML AND CGI PROGRAMMING

Charles J. Harris
Graduate Student
Department of Computer Science
State University of New York Institute of Technology

Abstract

As more machines are able to access data on the Internet, access to data through this medium will become more important. Among the reasons for this approach are platform-independence, ease of use, and minimal client requirements. In this paper, a project implementing a Web-based Interface to an image database will be discussed.

A WEB BROWSER DATABASE INTERFACE
USING HTML AND CGI PROGRAMMING

Charles J. Harris

Today, everyone from children to politicians seems to be talking about the World Wide Web (WWW) and the Internet. However, a few short years ago, the Internet was almost exclusively the domain of college students and computer programmers. As the Internet's user community evolved, so did the software that was used; from arcane, text-based, UNIX utilities grew today's network applications, with lucid graphical interfaces and intuitive controls. In fact, it is arguable that without the growth of the Web and the development of browsers such as NCSA Mosaic, the Internet would be a great deal less accessible than it is. The Web browsers enabled the user to access virtually any of the existing Internet services, such as Gopher and the File Transfer Protocol (FTP), through a single interface, greatly simplifying a system that was already extremely confusing. As the Internet has grown, it has developed around the Web, with each advance being incorporated into the Web's language, the HyperText Markup Language (HTML). Thus, the Web and its browsers act as a unifying and simplifying force on the Internet.

One might ask whether these other programs could benefit from the same type of interface. Work done at Rome Laboratory at Griffiss Air Force Base indicates that it can. A team there is developing an interface to the Image Product Archive (IPA) database based on Web programming and the Mosaic and Netscape browsers.

Web Browsers and the Common Gateway interface (CGI)

Though the vast majority of Web documents are static pages of hypertext, modern web browsers and servers also allow readers to view the output from applications running on the server machine. This is done through the Common Gateway Interface (CGI). When a browser requests the name of a CGI program or script, the server will run the application and send the output back to the requesting browser. Typically, the CGI program will output HTML code to build a page incorporating whatever information was requested.

The CGI application can be any program executable on the server machine. This means it can be a compiled binary program built from any language. Alternately, one can use a scripting language, such as the UNIX shell script commands or Perl. The server configuration files enforce certain security limitations upon the executed programs. Only certain directories are allowed to contain CGI programs, and it is assumed that administrators will police those directories and prevent their users from adding questionable programs to the server directories. In addition, some servers lock any spawned programs to a certain directory tree, prohibiting them any access to other, more sensitive areas of the file system.

There are two basic approaches to writing a CGI program. The first is to build a normal application which can be called from a CGI script. While this allows a single application to serve users on both the WWW and the actual system, often this approach limits what can be done, preventing either audience from exploiting the advantages of their system. In addition, this may require writing output to temporary files where they can be incorporated into the HTML file, adding overhead to the whole transaction. Alternately, an application can be designed to be used exclusively as a CGI program, handling any system tasks and formatting output directly into HTML. This generally results in tighter, cleaner, code with less overhead than the scripted approach. In addition, in systems where security is an issue, it makes it much more difficult for a hacker to use the Web server to break into the computer.

Input is handled primarily through a form-based interface. The user is presented with a page of input fields, including text entry fields, radio boxes, buttons, and check boxes. In addition, most browsers support an imagemap capability. This allows the designer to designate areas of an image as links to other HTML pages, making possible features such as button bars and active hypertext maps.

The main disadvantage of using the CGI to write programs is its static nature. Unlike a normal graphic application, the HTML page generated by the CGI program is static, and cannot be updated as the user makes changes to the form. The user must make all of his changes and then submit the form for any

action to be taken. As it can take some time to update the page, submitting the form after every change can quickly become a cumbersome operation. To help alleviate this problem, some modern browsers have included support for Sun Microsystems' Java programming language, which allows applications to perform virtually any function of a standard windowed interface. In addition to true Java, Netscape has added a scripted version called Javascript, permitting some aspects of the page to be determined dynamically. However, currently very few browsers support these extensions. Hence, if one wants to ensure compatibility with older browsers, these features are unavailable.

The basic method for coding a CGI program is to first create a form that allows the user to enter any relevant data. In the case of a program that does not accept input, this can be as simple as a single submission button or link, but typically the user must enter some data for the execution to proceed. Then, once the user has filled any necessary fields in the form, she clicks on the submission button, which has been keyed to the CGI program to interpret the form. The data from the form is sent to the CGI program by the server, and the program processes any actions that need to be taken. The CGI application finally creates its output HTML and sends the result back to the user's browser. Of course, this output can include more form structures, allowing the user to perform additional actions.

Advantages of Web-Based Applications

There are many advantages to CGI programs. Perhaps the biggest advantage is the platform independence that they permit. Virtually any computer can be used as a client to the server, as Web browsers are available for nearly every modern platform and operating system. This means, for instance, that the application can be accessed from whatever machine may be on an employee's desk, preventing both the cost of an expensive workstation and also the cost and time loss of training an employee to use a different computer so she can utilize a single application.

In addition, the Web browser and server typically contain all of the networking code required for the application. Thus, programmers can focus their efforts on the application code instead of writing network

code for every different environment on which the client may be run. Moreover, if security is an issue, secure clients and servers are available that encrypt transmissions automatically, reducing the chance that sensitive information can be intercepted on the network line. For sites on the Internet, the integral networking means that customers can access your application from nearly anywhere in the world as easily as they could if they were in the same room as the server.

Web-based programming also minimizes the maintenance of client machines. Nearly all of a CGI application is stored and executed on the server. The only components on the client machine are a browser and whatever plug-ins or file viewers are needed for the types of data to be accessed; generally, these programs will be on the machine anyway. Thus, upgrading the application is likely to consist entirely of changing a few files on the server, but leaving the client machines untouched; this is clearly preferable to upgrading the software on every client machine in a large installation.

The Web-based interface is typically intuitive and comfortable to new users. The Web was designed to provide an interface that everyone could understand, and this translates directly to the Web-based client. The controls on a browser are simple to operate, and the additional features implemented within the pages can easily be learned, since they build upon a clear model.

The IPA Web Browser Interface

The Image Product Archive database is used to store intelligence photographs along with related data such as sources, the date the image was acquired, target information about the subject, etc. In addition, other data formats may be stored, such as animations, sound files, and documents for various applications. The system, designed by GTE, is built on the Sybase database engine and runs only on Sun Microsystems Sparc and DEC Alpha processors. The system is provided with an X-Windows client for querying the database. This client is extremely complex; not only does it require several steps to retrieve a product, it assumes that the user is at least superficially familiar with the schema used in the database Also, since

the client will only run under X-Windows, it restricts the types of machines that can query the database to expensive UNIX workstations.

Clearly, there are several problems with such a client. A UNIX workstation is not trivial to use or administer. By the same token, these machines may well be too expensive to issue to everyone who needs access to the database. While X-Servers are available for PC's, allowing them to run remote X-Windows applications, this requires everyone to run their clients from a central UNIX server. This overloads resources in two ways; first, the clients themselves require memory and processor resources on the server, and second, the network can easily be overloaded as several clients attempt to update their displays on PC's at the same time. Even if sufficient computing and network resources are available, the significant investment required for user training must be considered. In addition, there are several other intelligence product databases in use by the users of IPA, such as the Imagery Dissemination and Exploitation (IDEX) and Imagery Exploitation Support System (IESS). Needless to say, the IPA client does not allow the user to access these databases as well, and the clients for each of them is somewhat different than the IPA client, forcing some users to remember three completely different interfaces.

The IPA Web Browser Interface developed at Rome Laboratories is an attempt to correct many of these shortcomings. The system utilizes the Mosaic and Netscape web browsers and a series of CGI programs on the IPA server to access the database and display the information on any computer capable of running either of the two supported browsers, including UNIX workstations, PC's, and Macintosh computers. In addition, the Interface is well suited to accessing the server over a modem dial-up line. Discounting the actual image transfer, the bandwidth requirements are considerably less than an X-Windows application that must be redrawn regularly.

The Web Browser Interface gives the user several query functions. The first, Browse, displays a list of keywords distilled from the database records. This allows a new user to easily retrieve relevant records even if she is not familiar with the total database. Less casual users might be interested in the Find utility.

This is a simple keyword search engine, looking for the input keyword in several of the more likely fields, such as title and target type. Users who are comfortable with the database schema will probably prefer the Query function, as it gives them access to the full capabilities of the Sybase database engine. The user can specify up to three search terms to find in different attributes; these terms can be combined with all of the expected boolean functions. Finally, a Geographic search feature is provided. The user specifies a point and a radius, and the interface will return all records that are geographically located within the circle. The current release of the interface requires the user to manually enter the latitude and longitude; the version under production now uses a series of maps to allow the user to click on map areas to define the circle. Also under development is a polygon search using the same interface, where the user will select each vertex of the polygon in turn and the interface will return the enveloped products. The next release will also give users the ability to save commonly performed queries and retrieve them at will.

The interface also allows users with the proper authorization to submit new products to the database. The user is prompted for relevant information about the new record, and it is copied to the server automatically via FTP. An indexing process runs regularly, incorporating the new record into the database and keyword indices.

While most queries to the system result in an image being immediately downloaded, some database entries refer to products stored off-line; e.g., on magnetic tape. In these cases, the request is submitted by the database and the Web Browser Interface will allow the user to periodically check the request status.

One of the design requirements of the system was the ability to work over relatively slow dial-up lines. For this reason, all of the functions of the interface can be run in either text or graphics mode. While the graphic buttons are small, they do slow the transfers down, and thus, are turned off in text mode. The mode can be toggled in nearly any screen. In addition, the user has the ability to view thumbnails of the images before she downloads the actual product, which can be several megabytes.

Security is maintained through the Web server. Before the user can access any of the database features, she must provide a valid username and password from the server machine. Rather than using the server's built-in authentication scheme, the program actually looks in the system's password file for valid combinations. This frees the server administrator from maintaining two distinct password files. If a server contains no sensitive data, it can also be configured as a Central Server. Here, no password is required, and clients can connect to multiple servers easily. However, most of the sites currently using the Interface are Protected servers, and require user authentication.

The IPA Web Browser Interface currently has a requirement to support versions of Mosaic later than 2.4 and Netscape Navigator versions starting at 2.0. Because of this, many of the fancier additions to HTML, including frames, tables, and Java, were not available for the project. However, an impressive interface was created using only basic CGI forms and C programs.

The success of the project can probably be best determined by looking at the system's use in the field. In slightly more than a year, the Web Browser Interface has moved from a basic proof of concept demonstrator to a system in use in dozens of sites around the world. At this point, the response has been overwhelmingly positive, to the point that the Web Browser Interface is currently being installed at the same time as the IPA database system itself.

Thus, it should be clear that Web-based programming presents a solution to many of the problems with modern applications. As features such as Java become more accepted, this trend can't help but accelerate. While the technology is not yet ready for real-time applications, it seems more than sufficient for many other situations, such as database access. This too can only improve as the Web grows to encompass the world.

References

Information about CGI programming can be found in a vast collection of books about HTML. One good source for this information is:

Savola, Tom. **Using HTML Special Edition**. Que Corporation, Indianapolis, 1995.

Several articles about writing interfaces specifically for databases are found in the September 1996 issue of Web Techniques.

INVESTIGATION OF SYNCHRONIZED MODE-LOCKED FIBER LASERS

Walter Kaechele
Graduate Student
Physics Department

Rensselaer Polytechnic Institute
110 8th Street
Troy, NY 12180

Final Report for:
Graduate Student Research Programs
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

September 1996

INVESTIGATION OF SYNCHRONIZED MODE-LOCKED FIBER LASERS

Walter Kaechele
Graduate Student
Physics Department
Rensselaer Polytechnic Institute

Abstract

A passively mode-locked fiber laser is synchronized to an actively mode-locked fiber ring laser. At injection powers of 1.5 mW stable synchronization is obtained, and several operation regimes are identified.

INVESTIGATION OF SYNCHRONIZED MODE-LOCKED FIBER LASERS

Walter Kaechele

I Introduction

With optical data streams achieving rates in excess of a terabit[1, 2], all-optical synchronization of fiber lasers for optical clock recovery has become an area of increasing interest. This method of synchronization utilizes an incoming optical signal or pulse train to drive a mode-locked laser, which in turn generates a pulse stream at the same rate as the incident signal. This signal can then be used for further optical processing of the data stream. Passively mode-locked lasers conveniently provide short pulses, but they suffer from low repetition rates and relatively large timing jitter. A possible solution would be to seed a series of passive lasers by an incoming signal so the individual lasers can run at their lower frequencies but combined would provide the data rates required. In addition to locking the laser to the master signal, timing jitter in passively mode-locked lasers has been shown to be reduced in some instances by injection locking[3, 4]. Therefore, a set of N synchronized lasers could provide the base for a distributed time division system with a common clock at the higher optical rates.

II Experiment

The experimental system required a passively mode-locked laser to be slaved to a series of injection pulses acting as the master signal. The source of the pulses or master laser was an actively mode-locked fiber ring laser shown if Fig. 1a. The output was in a definite polarization state defined by the polarization controller and the modulator. The Mach-Zehnder modulator also mode-locked the ring laser at harmonics of the fundamental cavity frequency. The cavity length was adjusted until the fundamental frequency of the master laser agreed with the slave laser's frequency of 3.5435 MHz within 100 Hz. By adjusting the polarization paddles within the cavity, pulsewidths of 2 ps were obtained. These pulses acted as the injection signal to the passively mode-locked laser. With the addition of the erbium doped fiber amplifier, the injection signal reached input powers up to 10 mW.

Most fiber laser synchronization schemes utilize a fiber ring laser configuration for the

passively mode-locked system[3, 5], however, the synchronization of a linear fiber cavity was recently reported by Min Jiang et al.[6]. We employed a standing wave erbium fiber laser in a Fabry-Perot configuration illustrated in Fig. 1b. The input and output was taken through a fiber Bragg grating which also served as a mirror for the cavity. The grating had a reflection of 50saturable absorber acted as the mode-locking element and as the second mirror for the laser. We eliminated the need for bulk optical mounts by epoxying the MQW saturable absorber to a cleaved fiber end, which allowed us to operate in an all-fiber system.

The master signal was introduced through the fiber Bragg grating via a 50/50 coupler. The output was taken through the unused port and sent to the diagnostic equipment. The MQW laser produced output in a completely random polarization state as none of the elements are polarization sensitive. Spectral and temporal results were observed and multiple stable operational regimes were recorded.

III Results

Mode-locked, synchronized operation was maintained at the fundamental frequency with average injection powers as low as 1.3 mW, which corresponds to pulse energies of 360 pJ. Below this level the master laser signal was insufficient to initiate synchronized behavior in the slave laser. No discernible difference in operation was observed in the synchronized output when the injected power was increased above this threshold. The synchronized output is shown in Fig. 2. The smaller pulses are from the ring laser with the larger being the slaved output from the MQW laser.

More important to stable operation was the position of the incident light on the surgace of the saturable absorber. By adjusting the fiber's location and distance from the surface of the MQW saturable absorber, differing regimes could be observed. Operating unsynchronized in a mode-locked state, the passive MQW laser produced 7 ps pulses with a spectral width of 0.45 nm. This gives a time-bandwidth product of 0.4, which is somewhat greater than the 0.31 value corresponding to hyperbolic secant pulses. The synchronized output produced three distinct stable pulse outputs with pulse widths of 10, 15, and 30 ps shown in Fig. 3. The optical spectrum remained stable in all three regimes with a bandwidth of 0.32 nm at FWHM.

IV Conclusions

Stable operation of a novel synchronization scheme has been achieved. As an application to current communication needs, pulse timing and amplitude stability are critical parameters in assessing potential system performance; time multiplexed versions in particular. RF noise analysis can provide an initial estimate of the jitter and amplitude fluxuations of all three laser systems; passive, active, and hybrid.

V Acknowledgements

The author would like to thank Dr. Kenneth Teegarden for help in construction of the MQW laser system. Thanks are also due to Dr. Joseph Haus and Reinhard Erdmann for useful discussions conserning the dynamics of fiber lasers and synchronization.

References

[1] E. Yamada, E. Yoshida, T. Kitoh, and M. Nakazawa, Elec. Lett. **31**, 1342 (1995).

[2] P. A. Morton, V. Mizrahi, G. T. Harvey, L. F. Mollenauer, T. Tanbunek, R. A. Logan, H. M. Presby, T. Erdogan, T. Sergent, and K. W. Wecht, IEEE Phot. Tech. Lett. **7**, 111 (1995).

[3] M. Margalit, M. Orenstein, and G. Eisenstein, Opt. Lett. **20**, 1877 (1995).

[4] J. K. Lucek, and K. Smith, Opt. Lett. **18**, 1226 (1993).

[5] K. Smith, and J. K. Lucek, Elec. Lett. **28**, 1814 (1992).

[6] Min Jiang, W. Sha, L. Rahman, B. C. Barnett, J. K. Andersen, M. N. Islam, and K. V. Reddy, Opt. Lett. **21**, 809 (1996).

Figures

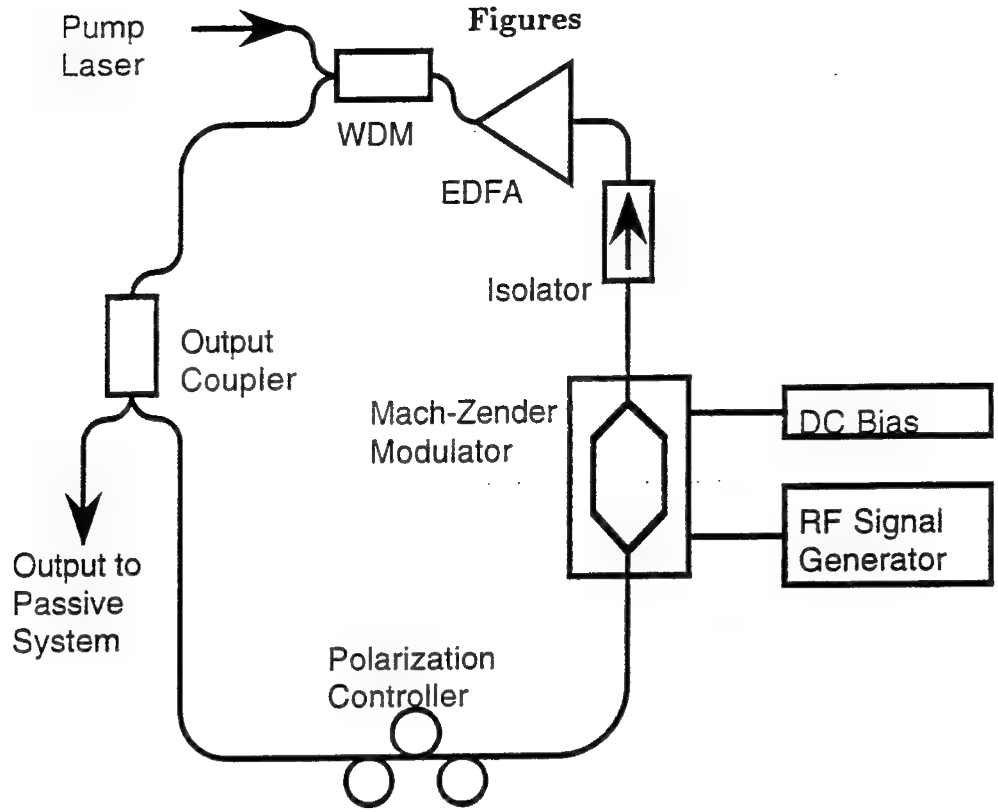


Fig. 1a. Layout of fiber ring laser used as master laser. WDM: Wavelength division multiplexer. EDFA: Erbium doped fiber laser. The output from this laser was used to synchronized the passive laser.

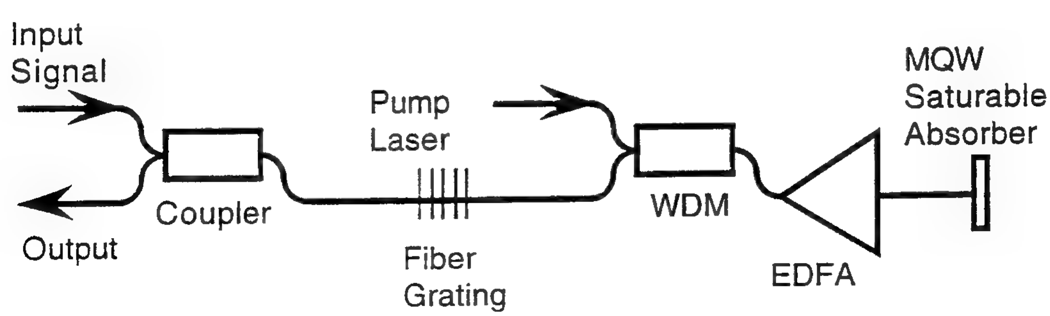


Fig. 1b. Layout of linear cavity employing multiple quantum well saturable absorber and fiber Bragg grating. The injected signal as well as the synchronized output are taken out through the fiber grating.

Figures

Fig. 2. Synchronized output as seen on a digitizing oscilloscope. The pulses are at an identical rate of 3.5435 MHz. The smaller pulses are from the ring laser, and the larger pulses are the synchronized output of the slave laser.

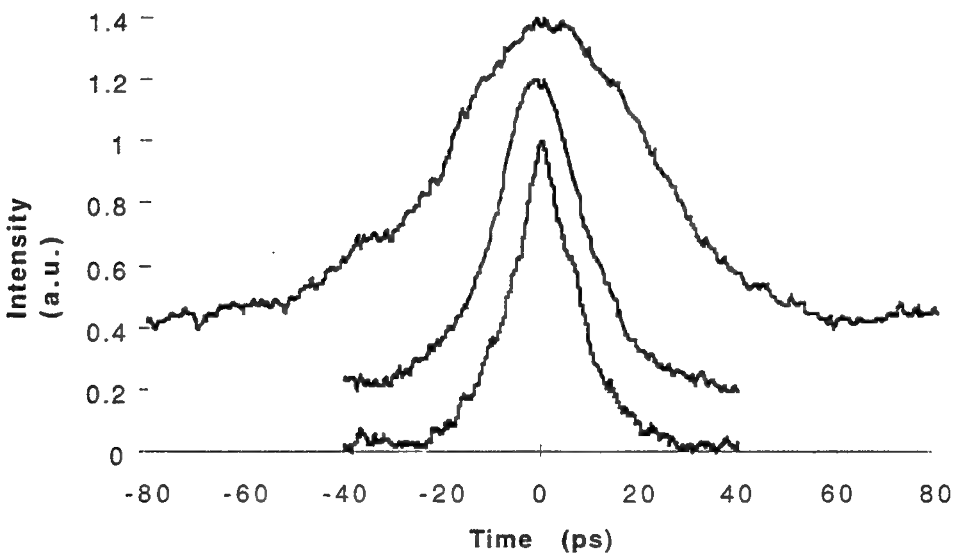


Fig. 3. Autocorrelation traces of synchronized output in the three stable operating regimes.The pulses have a common spectral bandwidth of 0.32 nm.

NON-GAUSSIAN CLUTTER MODELING BY SPHERICALLY INVARIANT RANDOM VECTORS

Andrew D. Keckler
Doctoral Candidate
Department of Electrical Engineering

Syracuse University
Department of Electrical Engineering and Computer Science
121 Link Hall
Syracuse, New York 13244

Final Report for
Summer Graduate Student Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Boling Air Force Base, DC

and

Rome Laboratory

September 1996

NON-GAUSSIAN CLUTTER MODELING BY SPHERICALLY INVARIANT RANDOM VECTORS

Andrew D. Keckler
Syracuse University, Department of Electrical Engineering and Computer Science
121 Link Hall, Syracuse, NY 13244

ABSTRACT

Conventional radar receivers are based on the assumption of Gaussian distributed clutter. As the resolution capabilities of radar systems improve, the validity of this assumption becomes questionable, and the clutter is often observed to be non-Gaussian. For example, the Weibull and K-distributions have been shown to approximate some experimentally measured non-Gaussian clutter data. In this environment, the detection performance of the Gaussian receiver may be significantly below that of the optimum non-Gaussian receiver. In order to obtain improved detection performance, it is necessary to characterize the correlated, non-Gaussian clutter samples. Spherically invariant random vectors (SIRV's) appear to be an appropriate model for the non-Gaussian clutter. A library of distributions that conform to the SIRV model is presented, as well as efficient techniques for simulating SIRV's. A technique for approximating an SIRV using a multivariate Gaussian-mixture distribution is also proposed.

Keywords: Spherically Invariant Random Vectors, non-Gaussian clutter, Gaussian-mixture

NON-GAUSSIAN CLUTTER MODELING BY SPHERICALLY INVARIANT RANDOM VECTORS

Andrew D. Keckler

1. INTRODUCTION

Historically, the clutter returns from radar systems were thought to be the sum of contributions from many scatterers, none of which were dominant. From the Central Limit Theorem, it was concluded that the Gaussian distribution was an appropriate statistical model for the clutter. However, as the resolution capabilities of radar systems have improved, the validity of this assumption has been brought into doubt. Experimental data for high-resolution radar systems has been observed to be non-Gaussian, particularly at low grazing angles[1-5]. Distributions such as the Weibull and K-distributions, which fall within a general class of multivariate distributions known as Spherically Invariant Random Vectors (SIRV's), have been shown to approximate some experimentally measured non-Gaussian data[3,4]. This is significant, since the performance of the Gaussian receiver in this environment may be well below that of the non-Gaussian receiver.

The improved performance of the non-Gaussian receiver may be understood by considering the probability density functions (PDF's) of such distributions as the Weibull and K-distribution. These PDF's typically have higher tails than the Gaussian PDF, which leads to more frequent occurrences of large clutter returns. These distributions are often described as "spiky", and this effect is illustrated in figure 1.1[6] for

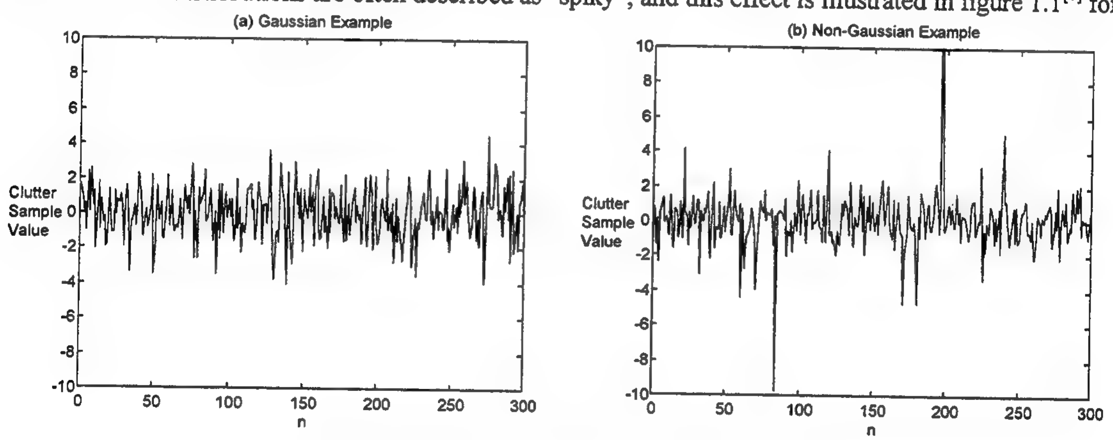


Figure 1.1: Time Sequence of "Spiky" Clutter Data.

distributions having the same variance. Conventional radar systems are plagued by the target-like spikes produced by the extended tails of the non-Gaussian distributions. In effect, this generates considerably more false alarms that require the threshold of the Gaussian receiver be raised in order to maintain the desired false-alarm rate for the non-Gaussian clutter, resulting in a reduction in the probability of detection. In contrast, the optimal non-Gaussian receiver is found to contain a nonlinearity which reduces large clutter spikes, allowing the threshold to be maintained at lower levels[6], which in turn provides increased target detection opportunities at the desired false alarm rate.

Knowledge of the clutter distribution is key to obtaining this improved performance. Optimal radar target detection requires the joint PDF of N pulse returns, which are collected from a particular range-azimuth cell during a coherent processing interval (CPI). A multivariate model is required, as pulse-to-pulse correlation may exist in the clutter returns of the received data vector. If the clutter samples are Gaussian, the form of the joint PDF is well known and is completely specified by the first and second order moments. However, knowledge of the first and second order moments is insufficient for the complete characterization of the non-Gaussian PDF. Generally it is difficult to specify multivariate non-Gaussian PDF's with independent control over the correlation properties. Specifically, it has been found to be difficult to independently control the shape of the distribution and its covariance matrix. Spherically invariant random vectors (SIRV's) are a class of non-Gaussian distributions that allow for the independent

control of the form of the distribution and its covariance matrix, making them a useful model for the interference. Closed form expressions exist for the multivariate PDF's of many types of SIRV's, and the class includes many distributions of interest, such as the Student-t, the Weibull, and the K-distributions. Furthermore, the Gaussian distribution is a member of this class.

Since clutter tends to be nonhomogeneous and nonstationary, different clutter models are necessary to characterize regions within the surveillance volume over time and space. In order to obtain improved detection performance, it is necessary to determine a suitable SIRV distribution for modeling the clutter in each region. Once a suitable model has been selected for a particular region, the appropriate processing can be employed. Therefore, a library of suitable SIRV models is required in order to implement the non-Gaussian processing. Furthermore, in order to evaluate the performance of the non-Gaussian algorithms, it is necessary to develop efficient techniques for simulating correlated non-Gaussian SIRV's. According to the Representation Theorem for SIRV's, an SIRV can be generated as the product of a univariate random variable and a Gaussian random vector with zero-mean, independent components. The generation of independent zero-mean Gaussian distributed random variables is well understood. Since SIRV's exhibit closure under linear transformations, the desired correlation can be imposed upon the SIRV by multiplying the random vector by the appropriate matrix.

The univariate multiplier is used to control the shape of the SIRV distribution, and its PDF, known as the characteristic PDF, may well have a mathematically complex form that does not lead to a closed form for its cumulative distribution function (CDF), or the inverse of the CDF. The inverse of the CDF is required to readily generate data directly from this distribution. Many approaches exist to find approximations for the inverse CDF, but the errors involved, especially in the tails of the distribution, may loom large in the evaluation of the non-linear, non-Gaussian receivers. If a function can be found that bounds the characteristic PDF from above, and for which an inverse CDF can be found, then the generalized rejection theorem can be used to generate data from the desired distribution without approximation error. The bound must closely fit the characteristic PDF, or a large percentage of the generated data points will be rejected, wasting time and depleting the available pool of random numbers.

The PDF of all SIRV's can be expressed in terms of a quadratic form $q(\mathbf{y})$ given by

$$q(\mathbf{y}) = (\mathbf{y}-\mathbf{b})^H \Sigma^{-1} (\mathbf{y}-\mathbf{b}) \tag{1.1}$$

where Σ is the covariance matrix and $\mathbf{b}$ is the mean vector of the SIRV $\mathbf{y}$. Often the problem of interest can be cast in terms of the quadratic form only. Since the quadratic form is always a scalar, it can be generated directly as a univariate random variable, and the above approach can be applied directly. This avoids the numerous problems associated with generating uncorrelated Gaussian random vectors. In some instances, the characteristic PDF of the SIRV is not known. In this case, an alternate technique for generating SIRV's using the envelope, which is directly related to the quadratic form given by equation (1.1), has been developed[7,8].

The characteristic PDF of SIRV's commonly considered, such as the Weibull or K-distributions, are continuous functions. This is not a requirement of the SIRV model, however. The discrete Gaussian-mixture SIRV[9-11] has the characteristic PDF,

$$f_s(S) = \sum_{k=1}^{K} w_k \delta(S - s_k), \quad \textit{for } w_k > 0 \quad \textit{and} \sum_{k=1}^{K} w_k = 1. \tag{1.2}$$

Through the proper choice of the parameters in equation (1.2), the discrete Gaussian-mixture SIRV can approximate many other types of SIRV's. This has many advantages, because the distribution of the discrete Gaussian-mixture SIRV is simple to evaluate, as is the distribution of its quadratic form. Random samples can easily be generated for the discrete Gaussian-mixture SIRV, and the form of the PDF for the discrete Gaussian-mixture SIRV leads to a parameterized receiver structure. This receiver can be used in place of the optimal receiver for any SIRV that the discrete Gaussian-mixture SIRV can approximate, without changing the structure of the receiver. Only the parameters need be adjusted.

In this paper, a library of known SIRV's is presented. Methods for generating random vectors consistent with these SIRV distributions are proposed and their performance is evaluated. The paper concludes by proposing a method for approximating SIRV's using Gaussian-mixtures.

2. SUMMARY OF THE SIRV MODEL

A brief review of the pertinent properties of the SIRV model is presented. The work of Yao [12] gives rise to a Representation Theorem for SIRV's that can be stated as follows:

Theorem 1: If a random vector is an SIRV, then there exists a non-negative random variable s such that the PDF of the random vector conditioned on s is a multivariate Gaussian PDF.

A spherically invariant random vector $\mathbf{x}$ with N zero-mean uncorrelated elements can therefore be represented by

$$\mathbf{x} = s\mathbf{z}, \tag{2.1}$$

where $\mathbf{z}$ is a Gaussian random vector with N zero-mean independent components, and s is an independent, non-negative random variable with PDF $f_s(S)$. This PDF uniquely determines the SIRV, and is called the characteristic PDF of the SIRV. It can be normalized such that $E(s^2) = 1$ without loss of generality [7].

A second property of SIRV's allows correlation to be introduced by a linear transformation, as stated in Theorem 2 [7]:

Theorem 2: If $\mathbf{x}$ is an SIRV with characteristic PDF $f_s(S)$, then

$$\mathbf{y} = [A]\mathbf{x} + \mathbf{b} \tag{2.2}$$

is also an SIRV with the same characteristic PDF, as long as the matrix [A] is nonsingular and $\mathbf{b}$ is a known vector having the same dimension as $\mathbf{x}$.

The transformed vector $\mathbf{y}$ will have a mean vector $\mathbf{b}$ and a covariance matrix $\Sigma = [A][A]^H$. The PDF of the SIRV is the joint PDF of its N components and is given by

$$f_y(\mathbf{Y}) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} h_N(Q(\mathbf{Y})) \tag{2.3}$$

where $Q(\mathbf{Y})$ is the quadratic form

$$Q(\mathbf{Y}) = (\mathbf{Y}\text{-}\mathbf{b})^H \Sigma^{-1} (\mathbf{Y}\text{-}\mathbf{b}) \tag{2.4}$$

and $h_N(Q(\mathbf{Y}))$ is a positive, real-valued, monotonic decreasing function given by

$$h_N(Q(\mathbf{Y})) = \int_0^\infty S^{-N} e^{-(\frac{Q(\mathbf{Y})}{2S^2})} f_s(S)\, dS \tag{2.5}$$

It can be seen that any N-dimensional SIRV is uniquely determined by its covariance matrix, mean vector and either its characteristic PDF or $h_N(Q(\mathbf{Y}))$. Furthermore, the PDF of the quadratic form $Q(\mathbf{Y})$ is given by

$$f_q(Q(\mathbf{Y})) = \frac{1}{2^{\frac{N}{2}}\Gamma(\frac{N}{2})} Q(\mathbf{Y})^{\frac{N}{2}-1} h_N(Q(\mathbf{Y})), \quad 0 \le Q(\mathbf{Y}) \tag{2.6}$$

This has an important implication. Since $h_N(Q(\mathbf{Y}))$ uniquely identifies each type of SIRV, this indicates that the multivariate problem can be reduced to an equivalent univariate problem in many instances. Since the quadratic form q(y) is a univariate random variable, it can be generated directly with the techniques proposed here.

3. PROBABILITY DENSITY FUNCTIONS FOR SIRV'S WITH UNKNOWN CHARACTERISTIC PDF'S

For some well known distributions, such as the Weibull and the Rician, the characteristic PDF is unknown. In addition, it may be difficult to evaluate the integral of equation (2.5) in closed form for some characteristic PDF's. An alternate technique has been developed by M. Rangaswamy[7,8] where the function $h_{2N}(Q(\mathbf{Y}))$ can be obtained directly from the first order (when N=2) envelope distribution. Generally this technique is limited to even order PDF's, but this does not present a problem as the data vector usually consists of complex quadrature components.

The function $h_2(Q(\mathbf{Y}))$ is directly related to the first order envelope PDF $f_r(R)$. As shown in equation (2.4), the quadratic form $Q(\mathbf{Y})$ is the square of the envelope for the whitened vector, and is related to the envelope PDF by

$$h_2(Q(\mathbf{Y})) = h_2\left(\frac{R^2}{\sigma^2}\right) = \frac{\sigma^2}{R} f_r(R), \qquad (3.1)$$

where σ^2 denotes the common variance of the in-phase and out-of-phase quadrature components.

Thus, equation (3.1) provides a method for obtaining the function $h_2(Q(\mathbf{Y}))$. From this, we can obtain $h_{2N}(Q(\mathbf{Y}))$ in a straight forward manner. This relationship is given by

$$h_{2N+2}(Q(\mathbf{Y})) = (-2)^N \frac{d^N h_2(Q)}{dQ^N}. \qquad (3.2)$$

Starting with $h_2(Q(\mathbf{Y}))$ and equation (2.3), all PDF's of even order can be generated. Similarly, all PDF's of odd order can be obtained from $h_1(Q(\mathbf{Y}))$ using

$$h_{2N+1}(Q(\mathbf{Y})) = (-2)^N \frac{d^N h_1(Q)}{dQ^N}. \qquad (3.3)$$

Since not all non-Gaussian envelope PDF's are admissible for characterization as SIRV's, the function $h_2(Q(\mathbf{Y}))$ and all of its derivatives must be examined to insure that they are positive, real-valued, monotonic decreasing functions.

4. ACCEPTANCE-REJECTION METHODS

A brief review of the rejection theorem is presented. The rejection theorem can be stated as:

Theorem 3: Let s be a random variable with density $f_s(S)$ and ϕ be any random variable with density $f_\phi(\Phi)$ such that $f_s(S)=0$ whenever $f_\phi(\Phi)=0$. Then let u be uniformly distributed on the interval (0,1). If ϕ and u are statistically independent and

$$\eta = \{U \le T(\Phi)\} \qquad (4.1)$$

where

$$T(\Phi) = \alpha f_s(\Phi)/f_\phi(\Phi) \le 1, \qquad (4.2)$$

then the rejection theorem states

$$f_{\phi|\eta}(\Phi|\eta) = f_s(\Phi). \qquad (4.3)$$

The density $f_\phi(\Phi)$ approximates $f_s(S)$ if the value

$$\alpha = \max_s f_s(S)/f_\phi(S) \qquad (4.4)$$

is a constant close to 1. If α equals 1, then f_s is identical to f_ϕ. In figure 4.1, the function αf_ϕ is seen to bound f_s in the sense that $\alpha f_\phi(S) \geq f_s(S)$ for all S in the support of s. It is desired to generate a random variable s

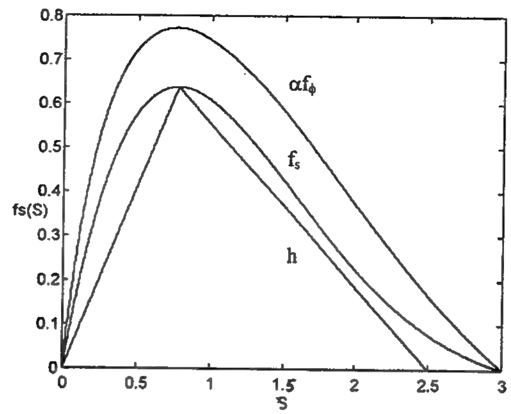


Figure 4.1: Illustration of the Acceptance-Rejection Method.

with density $f_s(S)$ from variates generated from $f_\phi(\Phi)$. This can be accomplished using following algorithm[2]:

STEP 1. Generate Φ from $f_\phi(\Phi)$ and compute $T(\Phi) = f_s(\Phi)/\alpha f_\phi(\Phi)$,
STEP 2. Generate U as a random variable distributed uniformly over the interval (0,1).
STEP 3. If $U > T(\Phi)$, reject Φ and return to STEP 1, else accept Φ as a variate from $f_s(S)$.

If $f_s(S)$ is a time-consuming function to evaluate, and there exists a function h such that $h(S) \leq f_s(S)$ for all S in the support of s, then a fast, preliminary test can be made, as can be seen if figure 4.1. The modified procedure becomes:

STEP 1. Generate Φ from $f_\phi(\Phi)$ and compute $T_h(\Phi) = h(\Phi)/\alpha f_\phi(\Phi)$,
STEP 2. Generate U as uniform (0,1),
STEP 3. If $U \leq T_h(\Phi)$, accept Φ as a variate from $f_s(S)$,
STEP 4. Else, compute $T(\Phi) = f_s(\Phi)/\alpha f_\phi(\Phi)$. If $U \leq T(\Phi)$, accept Φ as a variate from $f_s(S)$, else reject Φ and return to STEP 1.

This procedure has a geometric interpretation. A point (Φ, Y) is generated in the region bounded by $\alpha f_\phi(\Phi)$ and the Φ-axis with probability $1/\alpha$. If the point falls within the region bounded by $h(\Phi)$ and the Φ-axis, accept Φ immediately. If not, then if the point falls within the region bounded by $f_s(\Phi)$ and the Φ-axis, accept Φ. Otherwise reject Φ. The parameter α equals the area under the bound function, and the average efficiency of the acceptance-rejection algorithm is equal to $1/\alpha$.

5. GENERATION OF BOUNDS

In general, it is possible to partition a probability density function into M equal intervals such that horizontal line segments can be used to approximate the PDF, as shown in figure 5.1. It is assumed that the PDF is bounded and that the number of intervals is sufficient such that only one minima or maxima occurs in each interval. The equation for the bound function in this region is

$$\alpha f_\phi(S) = fu_k, \quad S_0 + (k-1)\Delta S \leq S < S_0 + k\Delta S, \; 1 \leq k \leq M, \tag{5.1}$$

where ΔS is the width of each interval and fu_k is the maximum value of $f_\phi(S)$ in the k^{th} interval.

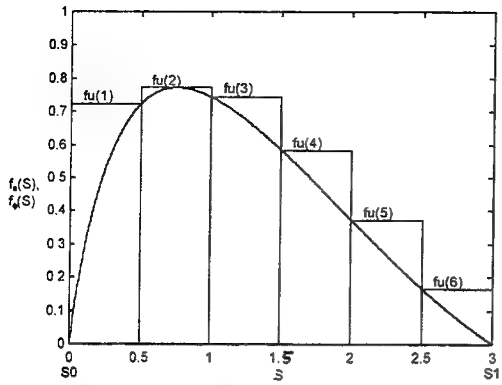


Figure 5.1: Piecewise Linear Bound of PDF.

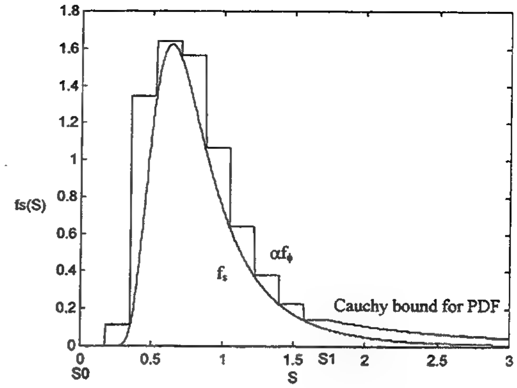


Figure 5.2: Cauchy Bound of PDF Tail.

Obviously, if the density has an infinite tail, the entire support cannot be segmented. The PDF is divided into a "body" and a "tail" portion at a point sufficiently far into the tail. This is illustrated in figure 5.2. Random variates can easily be generated for the Cauchy distribution, which is given by

$$f_x(X) = \frac{b}{\pi(b^2 + (X-\mu)^2)} \tag{5.2}$$

The Cauchy distribution is used to bound the tail portions of the PDF. It is assumed that the tail of the PDF $f_s(S)$ smoothly approaches zero at a rate faster than $1/S^2$. If this is not the case, a density function that approaches zero more slowly must be substituted.

The area under the bound in figure 5.2 can be easily calculated. Normalizing the bound by its area converts it to a probability density function for which the cumulative density function (CDF) is easily determined. This CDF is piecewise linear, except in the tail regions, and is easily invertible. The area encompassed by the bound of the tail, if the Cauchy PDF is used as the bound, is given by

$$F_x(X) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{X-\mu}{b}\right) . \tag{5.3}$$

Equation (5.3) is easily incorporated into the CDF of the bound, and is also easily invertible. To generate a variate from the bound PDF f_ϕ, a uniform random sample U is generated. Then

$$\Phi = F_\phi^{-1}(U) \tag{5.4}$$

will have the desired distribution. This provides an efficient and automatic technique for generating random variates for the bound required by the rejection algorithm of section 4.

6. GENERATION OF SIRV'S

In this section two procedures for generating SIRV's are presented. The first method applies when the characteristic PDF of the SIRV, $f_s(S)$, is known. The second procedure is applicable when the characteristic PDF is not known. According to the Representation Theorem, an SIRV is the product of a scalar random variable s and a Gaussian vector **z**. Thus, the simulation procedure for SIRV's is straightforward when the characteristic PDF is known, and is given by

STEP 1: Generate a sample of a zero-mean Gaussian random vector with unity covariance matrix,

STEP 2: Generate a sample of the random variable s from the characteristic PDF $f_s(S)$,

STEP 3: Calculate the product $\mathbf{X}=\mathbf{SZ}$ to generate the white zero-mean SIRV,
STEP 4: Perform the linear transformation $\mathbf{Y} = [A]\mathbf{X}+\mathbf{b}$ to obtain an SIRV with the desired mean and covariance matrix.

This is illustrated in figure 6.1. The covariance matrix Σ_x of the SIRV $\mathbf{x}$ is given by

$$\Sigma_x = E\{s^2\}\Sigma_z. \tag{6.1}$$

Usually, the random variable s can be normalized to have unit mean square value so that the SIRV has the same covariance matrix as the underlying Gaussian random vector. This can lead to difficulty in cases where the mean square value of the random variable s is infinite. This can also lead to problems with defining the quadratic form of equation (2.4). Alternatively, the quadratic form can be defined with the Gaussian covariance matrix Σ_z instead of the SIRV covariance matrix Σ_x.

When the characteristic PDF is unknown, an alternate technique for generating the SIRV's has been developed by M. Rangaswamy[7,8]. A white zero-mean SIRV can be expressed in generalized spherical coordinates as

$$\mathbf{X}_1 = R\cos(\phi_1), \tag{6.2}$$

$$\mathbf{X}_k = R\cos(\phi_k)\prod_{i=1}^{k-1}\sin(\phi_i), \quad (1 < k \leq N-2), \tag{6.3}$$

$$\mathbf{X}_{N-1} = R\cos(\theta)\prod_{i=1}^{N-2}\sin(\phi_i), \tag{6.4}$$

and

$$\mathbf{X}_N = R\sin(\theta)\prod_{i=1}^{N-2}\sin(\phi_i), \tag{6.5}$$

where $R\in(0,\infty)$, $\theta\in(0,2\pi)$, and $\phi_k\in(0,\pi)$ for $k=1,\ldots N-2$. A theorem of SIRV's states that the random variables R, θ, and ϕ_k are mutually independent. Furthermore, the distributions of the angles θ and ϕ_k are functionally independent of the white SIRV considered. Only the PDF of the envelope R varies from one SIRV to another. Since the multivariate Gaussian distribution is an SIRV, it can be used to generate samples from the angle distributions. It follows that

$$\frac{\mathbf{X}_k}{R} = \frac{\mathbf{Z}_k}{R_Z}, \; k = 1,2,\ldots N, \tag{6.6}$$

where R is the norm of the desired white SIRV and R_z is the norm of the zero mean white Gaussian vector. Therefore, the desired SIRV is obtained from

$$\mathbf{X} = \frac{R}{R_Z}\mathbf{Z}, \tag{6.7}$$

and leads to the following simulation procedure:

STEP 1: Generate a sample of a zero-mean Gaussian random vector with unity covariance matrix,
STEP 2: Compute the norm $R_Z=(\mathbf{Z}^T\mathbf{Z})^{1/2}$ of the sample vector $\mathbf{Z}$,
STEP 3: Generate a sample of the norm $R=Q^{1/2}$ from the PDF of the quadratic form of the SIRV,
STEP 4: Calculate a sample of the white SIRV $\mathbf{X}$ by computing equation (6.7),
STEP 5: Perform the linear transformation $\mathbf{Y} = [A]\mathbf{X}+\mathbf{b}$ to obtain an SIRV with the desired mean and covariance matrix.

This procedure is shown in figure 6.2. The generation of random samples from both the characteristic PDF $f_s(S)$ and the PDF of the quadratic form $f_q(Q)$ are not trivial. Generally it is not possible to evaluate the cumulative distribution function or its inverse. Thus, it is often necessary to generate samples using the rejection method as described in sections 4 and 5.

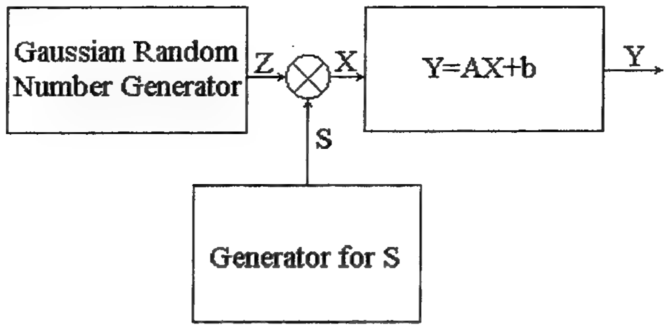


Figure 6.1: Generation of SIRV's with Known Characteristic PDF.

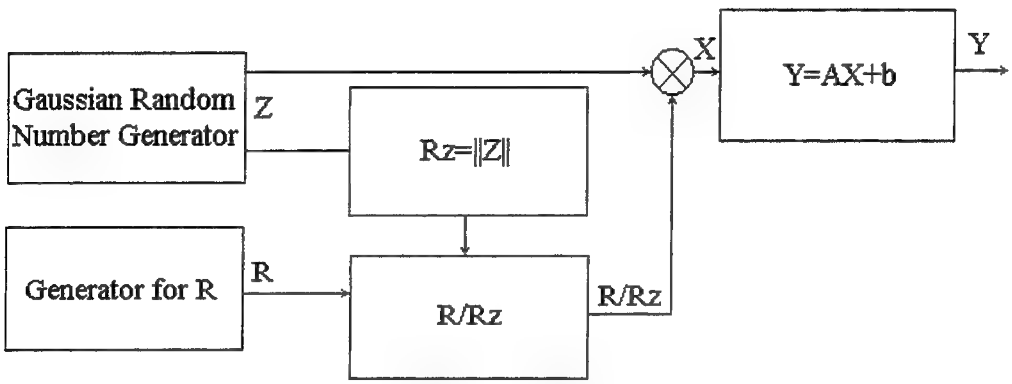


Figure 6.2: Generation of SIRV's with Unknown Characteristic PDF.

7. SIRV'S WITH KNOWN CHARACTERISTIC PDF

In this section results are presented for those SIRV's for which the characteristic PDF is known. Specifically, results are presented for the Laplace, Cauchy, Student-t, K-distribution, and Chi distributions. In addition, techniques for generating random samples from these distributions will also be presented.

7.1 The Laplace Distribution

The Laplace distribution is given by

$$f_x(X) = \frac{b}{2} e^{-b|X|}, \tag{7.1.1}$$

with scale parameter b. If the marginal distribution of the SIRV's quadrature components are distributed according to equation (7.1.1), then the characteristic PDF of the SIRV is given by

$$f_s(S) = b^2 S e^{-\frac{b^2 S^2}{2}} u(S), \tag{7.1.2}$$

where $u(S)$ is the unit step function. The mean square value of S is found to be

$$E\{S^2\} = \frac{2}{b^2}. \tag{7.1.3}$$

The function $h_N(Q)$ is given by

$$h_N(Q) = b^N\left(b\sqrt{Q}\right)^{1-\frac{N}{2}} K_{\frac{N}{2}-1}\left(b\sqrt{Q}\right)u(Q), \tag{7.1.4}$$

where $K_N(Q)$ is the N^{th} order modified bessel function of the second kind. The PDF of the quadratic form for the Laplace distribution is obtained from equation (2.6) as

$$f_q(Q) = \frac{b^2}{2\Gamma\left(\frac{N}{2}\right)}\left(\frac{b\sqrt{Q}}{2}\right)^{\frac{N}{2}-1} K_{\frac{N}{2}-1}\left(b\sqrt{Q}\right)u(Q), \tag{7.1.5}$$

where $\Gamma(n)$ is the eulero gamma function, which is defined by

$$\Gamma(n) = \int_0^\infty t^{n-1}e^{-t}dt,\ n > 0. \tag{7.1.6}$$

Random samples distributed according to the characteristic PDF given in equation (7.1.2) can be obtained by transforming exponentially distributed samples. The standard exponential distribution is given by

$$f_t(T) = e^{-T}u(T). \tag{7.1.7}$$

The random sample S is obtained from

$$S = \frac{\sqrt{2T}}{b}. \tag{7.1.8}$$

Figure 7.1.1 shows a histogram of 10,000 samples obtained from the characteristic PDF. Laplace distributed random vectors can be obtained by multiplying the samples obtained from the characteristic PDF by independent Gaussian vectors, and samples of the quadratic form can be obtained directly from the random vectors according to equation (1.1). Alternatively, samples of the quadratic form can be generated directly from the PDF given in equation (7.1.5) using rejection with the piecewise linear bound described in section 5. Figure 7.1.2 shows a histogram of 10,000 samples obtained from the PDF of the quadratic form, using the rejection method.

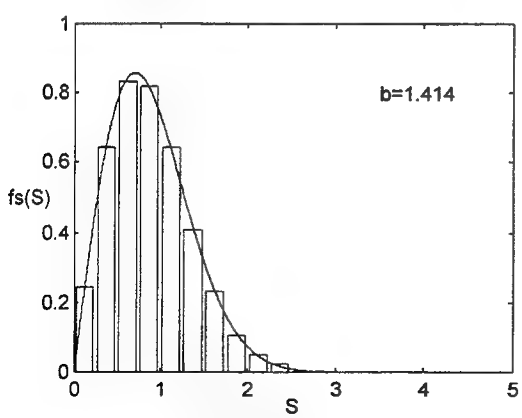


Figure 7.1.1: Analytic and Empirical Characteristic PDF's for the Laplace Distribution.

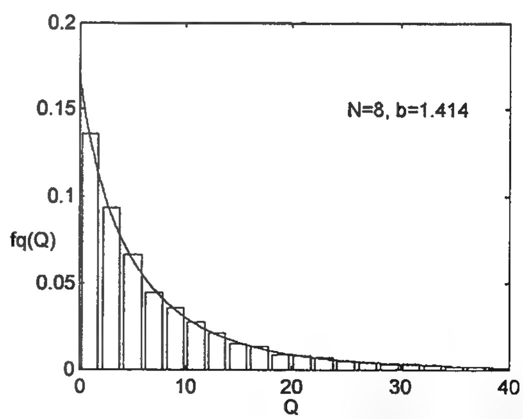


Figure 7.1.2: Analytic and Empirical PDF's for the Laplace Quadratic Form.

7.2 The Cauchy Distribution

The Cauchy distribution is given by

$$f_x(X) = \frac{1}{b\pi\left(1+\frac{X^2}{b^2}\right)}, \tag{7.2.1}$$

where b is the scale parameter. If the marginal distribution of the SIRV's quadrature components are distributed according to equation (7.2.1), then the characteristic PDF of the SIRV is given by

$$f_s(S) = \sqrt{\frac{2}{\pi}}\, bS^{-2} e^{-\frac{b^2}{2S^2}} u(S). \tag{7.2.2}$$

It should be noted that the mean square value of S is infinite. Because of this, the quadratic form of equation (1.1) is defined using the covariance matrix of the associated Gaussian distribution. With this in mind, the function $h_N(Q)$ becomes

$$h_N(Q) = \frac{2^{\frac{N}{2}} b\Gamma\left(\frac{N}{2}+\frac{1}{2}\right)}{\sqrt{\pi}\left(b^2+Q\right)^{\frac{N}{2}+\frac{1}{2}}} u(Q), \tag{7.2.3}$$

and the PDF of the quadratic form becomes

$$f_q(Q) = \frac{b\Gamma\left(\frac{N}{2}+\frac{1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{N}{2}\right)} \frac{Q^{\frac{N}{2}-1}}{\left(b^2+Q\right)^{\frac{N}{2}+\frac{1}{2}}} u(Q). \tag{7.2.4}$$

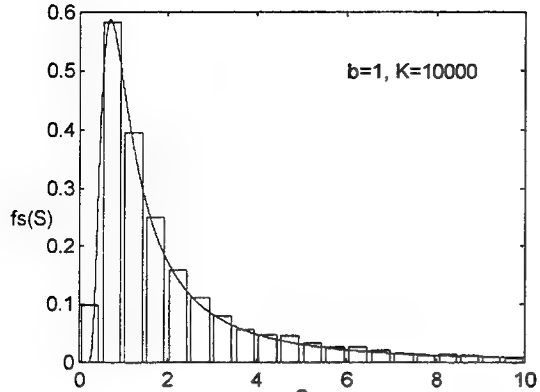
Random samples drawn from the characteristic PDF of equation (7.2.2) can be readily obtained using the rejection method with the piecewise linear bound described in section 5, or by transforming samples drawn from a gamma distribution with the shape parameter α set to ½. The PDF of the gamma distribution is given by

$$f_t(T) = \frac{T^{\alpha-1}}{\Gamma(\alpha)} e^{-T} u(T), \tag{7.2.5}$$

and samples of the random variable T can be readily generated from several standardized rejection algorithms, such as the IMSL subroutine RNGAM or the MATLAB function GAMRND. The samples of the random variable S are obtained from the transformation

$$S = \frac{b}{\sqrt{2T}}. \tag{7.2.6}$$

Figure 7.2.1 shows a histogram of 10,000 samples obtained for the characteristic PDF. Samples of the quadratic form are generated directly from the PDF given in equation (7.1.5) using the rejection method with a piecewise linear bound. Figure 7.2.2 shows a histogram of 10,000 samples distributed according to the quadratic form PDF.



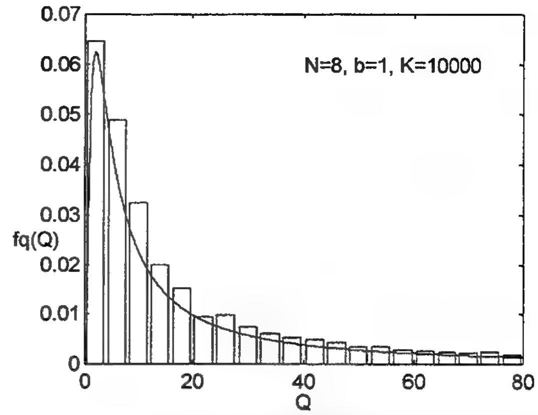


Figure 7.2.1: Analytic and Empirical Characteristic PDF's for the Cauchy Distribution.

Figure 7.2.2: Analytic and Empirical PDF's for the Cauchy Quadratic Form.

7.3 The Student-t Distribution

The Student-t distribution is given by

$$f_x(X) = \frac{\Gamma\left(\nu+\frac{1}{2}\right)}{b\sqrt{\pi}\,\Gamma(\nu)}\left(1+\frac{X^2}{b^2}\right)^{-\nu-\frac{1}{2}}, \tag{7.3.1}$$

and corresponds to the marginal distribution of the SIRV with shape parameter ν and scale parameter b. Note that the Cauchy SIRV is a special case of the Student-t SIRV, with ν equal to ½. The characteristic PDF of the Student-t SIRV is given as

$$f_s(S) = \frac{2b}{\Gamma(\nu)2^{\nu}} b^{2\nu-1} S^{-(2\nu+1)}\, e^{-\frac{b^2}{2S^2}} u(S), \tag{7.3.2}$$

and the mean square value of S is given by

$$E\{S^2\} = \frac{b^2}{2(\nu-1)}. \tag{7.3.3}$$

The function $h_N(Q)$ is given by

$$h_N(Q) = \frac{2^{\frac{N}{2}} b^{2\nu}\Gamma\left(\nu+\frac{N}{2}\right)}{\Gamma(\nu)\left(b^2+Q\right)^{\frac{N}{2}+\nu}} u(Q), \tag{7.3.4}$$

and the PDF of the quadratic form is

$$f_q(Q) = \frac{b^{2\nu}\,\Gamma(\nu+\frac{N}{2})}{\Gamma(\nu)\Gamma(\frac{N}{2})}\frac{Q^{\frac{N}{2}-1}}{(b^2+Q)^{\nu+\frac{N}{2}}} u(Q)\,. \tag{7.3.5}$$

Random samples for the characteristic PDF of the Student-t SIRV can be readily obtained using the same transformation as the Cauchy, with the shape parameter of the gamma distribution α set equal to ν. Figure 7.3.1 contains a histogram of 10,000 realizations of the random variable S for the characteristic PDF of the Student-t distribution. In the same manner as the Cauchy distribution, samples of the quadratic form are generated directly from the PDF given in equation (7.3.5) using the rejection method with a piecewise linear bound. Figure 7.3.2 shows a histogram of 10,000 samples obtained from the PDF of the quadratic form.

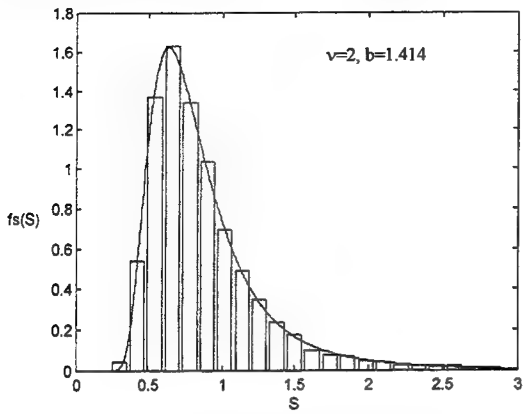


Figure 7.3.1: Analytic and Empirical Characteristic PDF's for the Student-t SIRV.

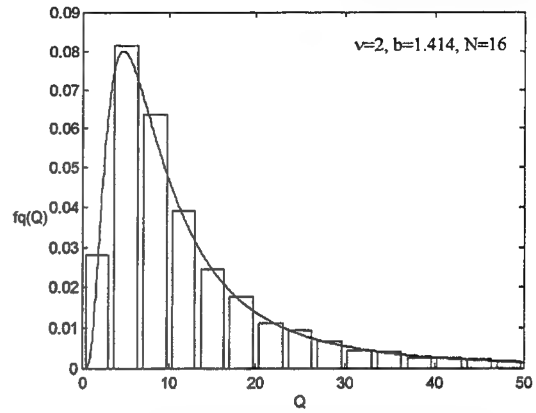


Figure 7.3.2: Analytic and Empirical PDF's for the Student-t Quadratic Form.

7.4 The K-Distribution

The K-distributed first order envelope PDF is given by

$$f_r(R) = \frac{2b}{\Gamma(\alpha)} \left(\frac{bR}{2} \right)^{\alpha} K_{\alpha-1}(bR) u(R), \tag{7.4.1}$$

where α is the shape parameter and b is the scale parameter, and where $K_N(Q)$ is the N^{th} order modified bessel function of the second kind. It should be noted that the K-distributed SIRV is defined by its first order envelope (when N=2), and not the marginal distribution of the quadrature components. The characteristic PDF is given by

$$f_s(S) = \frac{2b}{2^{\alpha}\Gamma(\alpha)} (bS)^{2\alpha-1} e^{-\frac{b^2S^2}{2}} u(S). \tag{7.4.2}$$

It should be noted that when α is equal to one, this characteristic PDF reduces to that of the Laplace SIRV, as given in equation (7.1.2). The mean square value of S is found to be

$$E\{S^2\} = \frac{2\alpha}{b^2}. \tag{7.4.3}$$

The function $h_N(Q)$ is given by

$$h_N(Q) = \frac{b^N}{\Gamma(\alpha)} \frac{\left(b\sqrt{Q}\right)^{\alpha-\frac{N}{2}}}{2^{\alpha-1}} K_{\frac{N}{2}-\alpha}\left(b\sqrt{Q}\right) u(Q), \tag{7.4.4}$$

and the PDF of the quadratic form is

$$f_q(Q) = \frac{b^2}{2\Gamma\left(\frac{N}{2}\right)\Gamma(\alpha)} \left(\frac{b\sqrt{Q}}{2}\right)^{\frac{N}{2}+\alpha-2} K_{\frac{N}{2}-\alpha}\left(b\sqrt{Q}\right)u(Q). \qquad (7.4.5)$$

Random samples can be generated for the characteristic PDF of equation (7.4.2) by transforming samples drawn from a gamma distribution with the same shape parameter. The samples of the random variable S are obtained from the transformation

$$S = \frac{\sqrt{2T}}{b}. \qquad (7.4.6)$$

Figure 7.4.1 shows a histogram of 10,000 samples obtained from the characteristic PDF of equation (7.4.2). Again, samples of the quadratic form are generated directly from the PDF given in equation (7.4.5) using the rejection method with a piecewise linear bound. Figure 7.4.2 shows a histogram of 10,000 samples obtained from the PDF of the quadratic form.

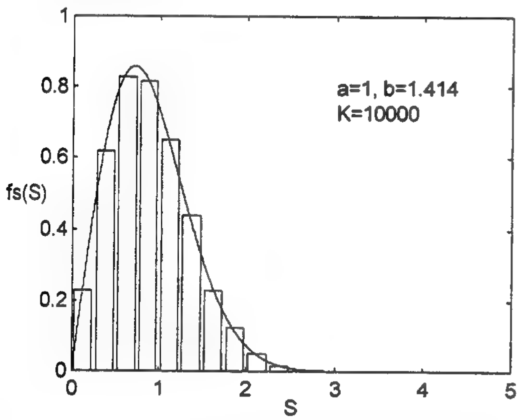


Figure 7.4.1: Analytic and Empirical Characteristic PDF's for the K-distributed SIRV.

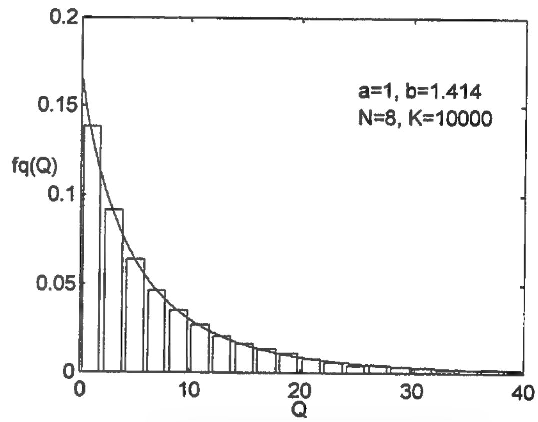


Figure 7.4.2: Analytic and Empirical PDF's for the K-distribution Quadratic Form.

7.5 The Chi Distribution

The Chi distributed first order envelope PDF is given by

$$f_r(R) = \frac{2b}{\Gamma(\nu)}(bR)^{2\nu-1} e^{-b^2R^2} u(R), \qquad (7.5.1)$$

where ν is the shape parameter and b is the scale parameter. Recent work by D.L. Stadelman[11] has shown that the characteristic PDF of the Chi envelope SIRV is given by

$$f_s(S) = \frac{2^{\nu+1} b^{2\nu}}{\Gamma(\nu)\Gamma(1-\nu)} \frac{S^{2\nu-1}}{\left(1-2b^2S^2\right)^{\nu}} u(S) u\left(\frac{1}{b\sqrt{2}} - S\right). \qquad (7.5.2)$$

An interesting feature of equation (7.5.2) that should be noted is that the characteristic PDF of equation (7.5.2) is non-zero for only a finite interval. The mean square value of S is found to be

$$E\{S^2\} = \frac{\nu}{2b^2}. \qquad (7.5.3)$$

The function $h_N(Q)$ is given by

$$h_N(Q) = \frac{(2)^{\frac{N}{2}} b^{2\nu}}{\Gamma(\upsilon)} \sum_{k=1}^{\frac{N}{2}} \binom{\frac{N}{2}-1}{k-1} Q^{\nu-k} b^{N-2k} \frac{\Gamma(k-\nu)}{\Gamma(1-\nu)} e^{-b^2 Q} u(Q), \tag{7.5.4}$$

and the PDF of the quadratic form is

$$f_q(Q) = \frac{1}{\Gamma(\upsilon)\Gamma\left(\frac{N}{2}\right)} \sum_{k=1}^{\frac{N}{2}} \binom{\frac{N}{2}-1}{k-1} Q^{\frac{N}{2}+\nu-k-1} b^{N+2\nu-2k} \frac{\Gamma(k-\nu)}{\Gamma(1-\nu)} e^{-b^2 Q} u(Q). \tag{7.5.5}$$

The characteristic PDF of equation (7.5.2) is related by a simple transformation to the beta distribution, and a beta distributed random sample can be generated from two gamma distributed random samples. Direct generation of random samples from the quadratic form is more difficult. For values of the shape parameter $\nu < .5$, the PDF of the quadratic form becomes infinite near the origin. This can be avoided by generating the transform of the quadratic form given by

$$Q = T^{\frac{2}{\nu}}. \tag{7.5.6}$$

The PDF of the transformed random variable t is given by

$$f_t(T) = \frac{2}{\nu\Gamma(\upsilon)\Gamma\left(\frac{N}{2}\right)} \sum_{k=1}^{\frac{N}{2}} \binom{\frac{N}{2}-1}{k-1} T^{\frac{2}{\nu}\left(\frac{N}{2}-k\right)+1} b^{N+2\nu-2k} \frac{\Gamma(k-\nu)}{\Gamma(1-\nu)} e^{-b^2 T^{\frac{2}{\nu}}} u(T), \tag{7.5.7}$$

and is bounded for all allowable values of the shape parameter ν. Figure 7.5.1 shows a histogram of 10,000 samples obtained for the PDF of the quadratic form, using the transformation specified in equation (7.5.6).

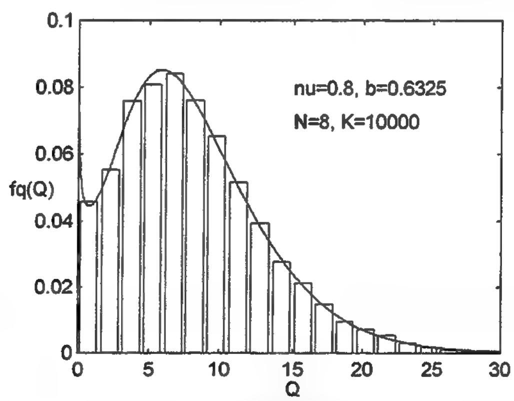


Figure 7.4.1: Analytic and Empirical PDF for the Chi Quadratic Form.

8. SIRV'S WITH UNKNOWN CHARACTERISTIC PDF

Working with the first order envelope (when N=2) and equations (3.2) and (3.3), it is possible to derive the probability density functions of an SIRV without knowledge of the characteristic PDF, as well as the distribution of the SIRV's quadratic form. M. Rangaswamy[7,8] has developed several such distributions, including the Weibull, Generalized Rayleigh, Generalized Gamma, and the Rician. In general, the form of these distributions are very complicated, and they are more difficult to work with.

Since the Weibull distribution has been shown to approximate some experimentally measured non-Gaussian data[4,5], great interest has been generated in it by the radar community. Therefore, results for the Weibull distribution will be presented here. The first order envelope (when N=2) of the Weibull SIRV has the distribution

$$f_r(R) = abR^{b-1}e^{-aR^b}u(R) \tag{8.1}$$

where a is the scale parameter and b is the shape parameter. The function $h_N(Q)$ is given by

$$h_N(Q) = (-2)^{\frac{N}{2}} e^{-AQ^{\frac{b}{2}}} \sum_{k=1}^{N} B_k \frac{A^k}{k!} Q^{\frac{kb}{2}-\frac{N}{2}} u(Q) \tag{8.2}$$

for 0<b<2, where

$$A = a\sigma^b, \tag{8.3}$$

$$B_k = \sum_{m=1}^{k} (-1)^m \binom{k}{m} \prod_{i=0}^{M-1} \left(\frac{mb}{2} - i\right), \tag{8.4}$$

and σ^2 is the common variance of the in-phase and out-of-phase quadrature components. The PDF of the quadratic form is given by

$$f_q(Q) = \frac{(-1)^{\frac{N}{2}}}{\Gamma(\frac{N}{2})} e^{-AQ^{\frac{b}{2}}} \sum_{k=1}^{\frac{N}{2}} B_k \frac{A^k}{k!} Q^{\frac{kb}{2}-1} u(Q) . \tag{8.5}$$

As in the case of the Chi SIRV, equation (8.5) becomes unbounded when the shape parameter b becomes less than 1. This causes problems for the rejection algorithm; and, since the characteristic PDF is unknown, it is necessary to generate Q directly. Using the transform

$$T = Q^{\frac{b}{2}}, \tag{8.6}$$

the probability density function

$$f_t(T) = \frac{2(-1)^{\frac{N}{2}}}{b\Gamma(\frac{N}{2})} e^{-AT} \sum_{k=1}^{\frac{N}{2}} B_k \frac{A^k}{k!} T^{k-1} u(T) \tag{8.7}$$

is obtained. This probability density function is bounded for all allowable values of the shape parameter b. Thus, equation (8.7) can be used to obtain random samples of the transform random variable T using the piecewise linear bound described in section 5. Then Q can be obtained from T using equation (8.6). Histograms of 10000 samples for the transform and the quadratic form of the Weibull distribution are shown in figures 8.1 and 8.2 respectively.

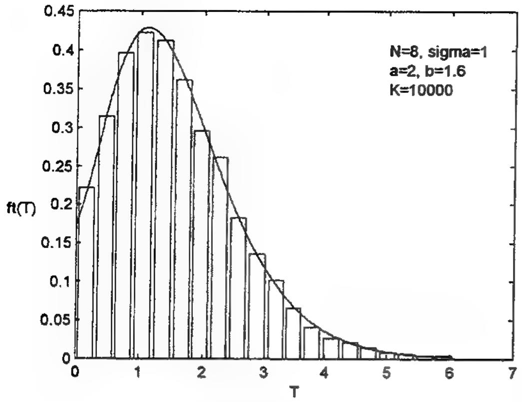


Figure 8.1: Analytic and Empirical PDF's for the Transformed Weibull Quadratic Form.

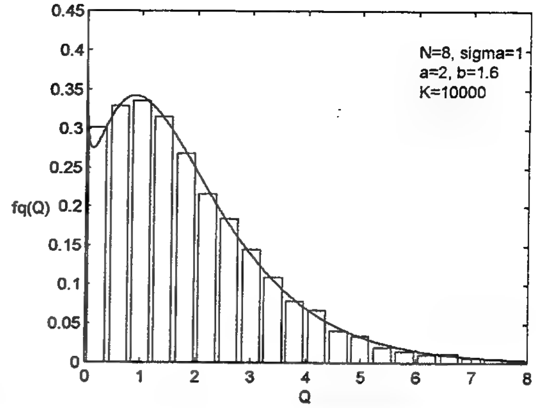


Figure 8.2: Analytic and Empirical PDF's for the Weibull Quadratic Form.

9. THE DISCRETE GAUSSIAN MIXTURE SIRV

The discrete Gaussian mixture SIRV has special significance in that its PDF is a finite weighted sum of Gaussian PDF's. This structure leads to some useful implications with respect to evaluating the Gaussian mixture PDF and the PDF of its quadratic form. It also leads to a simple technique for generating random samples. Through the proper choice of the parameters for the Gaussian mixture, it may be used to approximate many other SIRV's which may not be as readily evaluated.

The discrete Gaussian mixture SIRV has the characteristic PDF,

$$f_s(S) = \sum_{k=1}^{K} w_k \delta(S - s_k), \quad \textit{for } w_k > 0 \quad \textit{and} \sum_{k=1}^{K} w_k = 1, \tag{9.1}$$

and the PDF of the quadratic form of the discrete Gaussian mixture SIRV is

$$f_q(Q) = \frac{1}{2^{\frac{N}{2}} \Gamma(\frac{N}{2})} Q^{\frac{N}{2}-1} \sum_{k=1}^{K} w_k \, s_k^{-N} \, e^{-\frac{Q}{2s_k^2}}. \tag{9.2}$$

Consider again the PDF for the quadratic form of the Student-t SIRV given in equation (7.3.5), with N=16, b=1.414, and ν=2. For example, if the appropriate weights are selected, the discrete Gaussian mixture can be used to approximate the Student-t SIRV. The approximation achieved is shown in figure 9.1, and is a good fit for the body of the distribution. Random samples can be generated for the Gaussian mixture SIRV by simply generating a variate from the discrete distribution given in equation (9.1) and multiplying it by a Gaussian random vector. The discrete variate can be obtained by determining which bin a uniformly distributed variate falls into, and choosing the associated value of s_k. Likewise, data can be generated for the quadratic form of the Gaussian mixture by using the selected value of s_k as the variance of a gamma distributed random variable. A histogram of 10,000 random samples for the Gaussian mixture quadratic form specified in figure 9.1 is compared to the PDF of the Student-t quadratic form in figure 9.2.

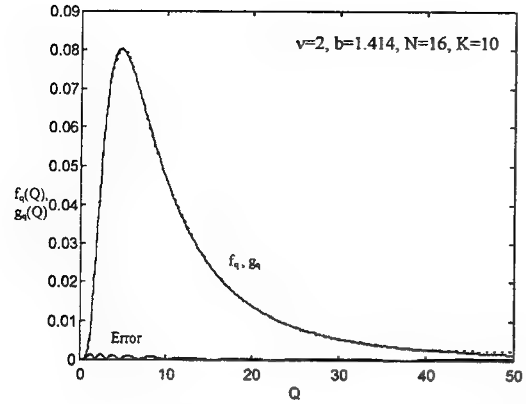


Figure 9.1: Comparison of Student-t Quadratic Form and the Gaussian Mixture Approximation.

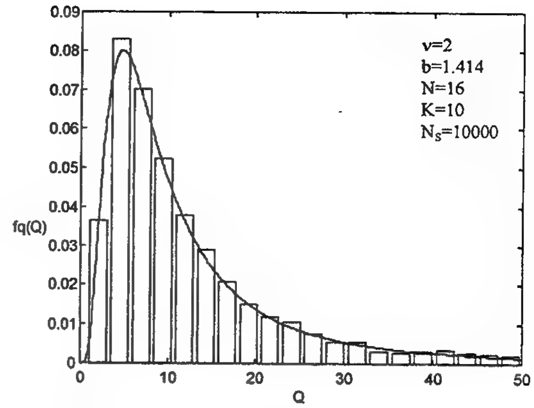


Figure 9.2: Comparison of Student-t Quadratic Form vs. Histogram of Gaussian Mixture Data.

10. SUMMARY

The SIRV model is found to be useful for modeling the non-Gaussian clutter problem. Significant improvement in detection performance has been demonstrated in the literature for the optimal non-Gaussian receiver over conventional receiver designs when the disturbance can be modeled as SIRV distributed clutter. In order to achieve this improved detection performance, a library of statistical models is required in order to approximate the interference. While this is necessary, it is not sufficient. Efficient techniques for monitoring the environment and selecting the appropriate non-Gaussian model are also required.

The use of the SIRV model has led to its own set of problems, however. For simulations, it is necessary that an efficient method of generating SIRV random samples be available. Through the use of a piecewise linear bound, an efficient, with respect to the number of points rejected, acceptance-rejection algorithm can be developed for any sufficiently smooth, bounded distribution. Furthermore, this procedure can easily be automated. The characteristic PDF's of many SIRV's fall within the above category. This can be combined with a Gaussian random number generator to produce multivariate distributions with the desired form and covariance matrix. Furthermore, if the application allows, the quadratic form of the SIRV can be generated directly with this method, avoiding the problems associated with generating independent random vectors. In addition, the Gaussian mixture SIRV can be used to approximate many other SIRV's. This leads to simple approximations for the PDF of a SIRV, and for the PDF of its quadratic form. A simple and highly efficient scheme exists for generating random samples for the Gaussian mixture SIRV, and thus random samples for many other SIRV's can be easily generated using the Gaussian mixture approximation. Furthermore, many other previously unknown SIRV's can be approximated and investigated using the Gaussian mixture SIRV, and the Gaussian mixture SIRV leads to a sub-optimal parameterized receiver which has the same form for all SIRV's.

11. ACKNOWLEDGMENTS

This work was supported by Rome Laboratory, U.S. Air Force through the Air Force Office of Scientific Research, 1996 Summer Graduate Student Research Program. Many of the SIRV models presented in this paper were built upon work first developed by M. Rangaswamy, D.D. Weiner and A. Ozturk in their paper "Computer Generation of Correlated Non-Gaussian Clutter for Radar Signal Detection". Finally, the author would like to thank Dr. Donald Weiner, without whose invaluable advice and experience, no progress would have been made.

12. REFERENCES

1. K.D. Ward, C.J. Baker, and S. Watts, *Maritime Surveillance Radar Part I: Radar Scattering from the Ocean Surface,* IEE Proceedings-F, vol. 137, pp 51-62, April, 1990.

2. C.J. Baker, *K-Distributed Coherent Sea Clutter*, IEE-Proceedings-F, vol. 138, no. 2, pp 89-92, April 1991.

3. J.K. Jao, *Amplitude Distribution of Composite Terrain Radar Clutter and the K-Distribution*, IEEE Transactions on Antennas and Propagation, vol. AP-32, no. 10, pp 1049-1062, October, 1984.

4. M. Sekine, S. Ohtani, T. Musha, T. Irabu, E. Kiuchi, T. Hagisawa, and Y. Tomita, *Weibull Distributed Ground Clutter,* IEEE Transactions on Aerospace and Electronic Systems, vol. AES-17, no. 4, pp. 596-598, July, 1981.

5. A. Farina, A. Russo, F. Scannapieco, and S. Barbarosa, *Theory of Radar Detection in Coherent Weibull Clutter*, IEE Proceedings-F, vol. 134, no. 2, pp 174-190, April 1987.

6. D.L. Stadelman and D.D. Weiner, *Detection of Weak Signals with Random Parameters in Non-Gaussian Clutter*, 1995 SPIE International Symposium on Optical Science, Engineering and Instrumentation, San Diego, CA, July, 1995.

7. M. Rangaswamy, D. Weiner, and A. Ozturk, *Non-Gaussian Random Vector Identification Using Spherically Invariant Random Processes*, IEEE Transactions on Aerospace and Electronic Systems, vol. 29, pp 111-124.

8. M. Rangaswamy, D. Weiner, and A. Ozturk, *Computer Generation of Correlated Non-Gaussian Radar Clutter*, IEEE Transactions on Aerospace and Electronic Systems, vol. 31, pp 106-116, January, 1995.

9. A.D. Keckler, *Generation and Approximation of Spherically Invariant Random Vectors*, 1995 AFOSR Final Report, Rome Laboratory, Rome, NY, September, 1995.

10. A. Keckler, Ph.D. dissertation in progress, Syracuse University, Syracuse, NY, 1996.

11. D. Stadelman, Ph.D. dissertation in progress, Syracuse University, Syracuse, NY, 1996.

12. K. Yao, *A Representation Theorem and Its Applications to Spherically Invariant Random Processes*, IEEE Transactions on Information Theory, vol IT-19, pp. 600-608, 1973.

13. D.E. Knuth, <u>The Art of Computer Programming Vol. 2: Seminumerical Methods</u>, Addison-Wesley, Reading, MA, 1981.

14. Paul Bratley, Bennett L. Fox, and Linus E. Schrage, <u>A Guide to Simulation</u>, Springer-Verlag, New York, NY, 1987.

15. W.H. Press, et. al., <u>Numerical Recipes in C: The Art of Scientific Computing</u>, Cambridge University Press, New York, NY, 1992.

16. K.J. Sangston and K.R. Gerlach, *Coherent Detection of Radar Targets in a Non-Gaussian Background*, IEEE Transactions on Aerospace and Electronic Systems, vol. 30, pp 330-340, April, 1994.

AN OVERVIEW OF THE SCHEDULING PROBLEM

Elizabeth I. Leonard
Ph.D. Candidate
Department of Computer Science

The Johns Hopkins University
3400 N. Charles Street
Baltimore, MD 21218

Final Report for:
Graduate Student Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

September 1996

AN OVERVIEW OF THE SCHEDULING PROBLEM

Elizabeth I. Leonard
Ph.D. Candidate
Computer Science Department
Johns Hopkins University

Abstract

In this paper we present an overview of the scheduling problem and the related constraint satisfaction problem. Because these problems are known to be NP-complete, good heuristics are necessary to efficiently search for solutions. One such heuristic, constraint propagation, involves maintaining the consistency of the domains of uninstantiated variables during the search for solutions. We review three techniques for maintaining consistency of variables while solving binary constraint satisfaction problems. We also consider the issue of benchmarks for scheduling problems and examine in depth a set of benchmarks for linear programming, the Kennington problems, which could be used to test scheduling algorithms.

An Overview of the Scheduling Problem *

Elizabeth I. Leonard

1 Introduction

Scheduling is defined in [Bak74] as the allocation of resources to tasks over time. Such problems occur frequently in production, transportation, and project management settings. This class of problems has been studied by researchers in the fields of operations research (OR) and artificial intelligence (AI). Because most scheduling problems are NP-complete, the optimization methods of OR are effective only for small problems. Search methods with good heuristics provide a way to find good, and sometimes optimal, solutions for more complicated problems.

Scheduling problems are instances of the more general constraint satisfaction problem. The tree search and consistency checking techniques used in solving constraint satisfaction problems can be applied to scheduling problems. The consistency checking techniques help to reduce the search space and to detect dead ends earlier in the search tree. Search and consistency checking techniques like those examined in this paper have been implemented in tools such as ILOG SOLVER [1].

Implementations of scheduling algorithms can be evaluated empirically by testing them on known benchmarks. The implementation's performance can be compared to that of other algorithms on the same standard tests. We review the results for some classic benchmarks for the most studied scheduling problem, the job shop scheduling problem. We also examine in depth a set of benchmarks for linear programming, the Kennington problems of [CHK+90], which could be used to test scheduling algorithms.

The rest of this paper is organized as follows. In Section 2 we review scheduling problems and the relationship to constraint satisfaction problems. We also examine the consistency checking techniques used in solving constraint satisfaction problems. In Section 3 we discuss some of the benchmarks available for testing scheduling problems. Finally, in Section 4 we end with some conclusions.

*Special thanks to Carla Gomes and Karen Alguire for helpful conversations and comments

[1] ILOG SOLVER is a product of ILOG Inc., 2005 Landings Dr., Mountain View, CA 94043

2 Scheduling Problems

Scheduling problems typically involve a set of resources (sometimes called machines) and a set of tasks (also called activities or operations). For each task, the machines on which it can be performed and the amount of time required for execution are specified. So are any precedence relationships which exist between tasks. There may also be deadlines and ready times given for the tasks. The problem is to allocate resources to the tasks over time without violating any of these constraints. Frequently the goal is to find the solution which optimizes some criterion. Usually this optimization problem is the minimization of the makespan, the total amount of time required to complete all the tasks.

The best known scheduling problem is the job shop scheduling problem (JSSP) which has been studied in [BPN95a, BPN95b, CL96, CL95, CL94]. In this problem there are m resources, each of which can process only one task at a time. Tasks are executed nonpre-emptively, i.e., once a task begins execution it finishes without interruption. The tasks are organized into n jobs, each composed of m tasks which are to be completed in a given order. It is assumed that no two tasks in the same job require the same resource. The durations of the tasks are also assumed to be known.

A variation of the JSSP is the multiple capacitated job shop scheduling problem of [NA96]. This differs from the standard JSSP in that resources are capable of processing multiple tasks simultaneously. Each resource has a specified capacity. Each task also has a size indicating how much of the resource it requires.

Another scheduling problem is the resource allocation problem [GH96]. There are n tasks which need to be allocated to m resources. The resources are identical in the sense that they are all capable of processing any of the tasks. In the specification some tasks may be excluded from executing on certain machines. This can be combined with the JSSP to give a job shop scheduling problem with multiple instances of resources as in [GTT94].

Scheduling problems can be thought of as a set of variables and constraints on those variables. The variables are the start times assigned to each of the tasks and the resource assigned to process the task. Constraints include the precedence relationships between tasks, capacity constraints for resources, and the resource requirements of the tasks. Thought of in this way, scheduling problems are instances of the constraint satisfaction problem.

2.1 The Constraint Satisfaction Problem and Consistency Techniques

Definition 2.1 *A* constraint satisfaction problem *P consists of a set $X = \{x_1, ..., x_n\}$ of variables each with a domain $D(x_i)$ and a set $C = \{c_1, ..., c_m\}$ of constraints over those variables where $c_i : D(x_1) \times \cdots \times D(x_n) \rightarrow \{true, false\}$. A solution to the problem P is an assignment $a \in D(v_1) \times \cdots \times D(v_n)$ such that*

Let y be the variable just instantiated with the value v.
Let $X = \{x_1, ..., x_n\}$ be the set of uninstantiated variables.
  For $i = 1$ to n
    if there is a constraint between x_i and y
      then for each $v' \in D(x_i)$
        if $x_i = v'$ and $y = v$ is inconsistent
          then $D(x_i) \leftarrow D(x_i) - v'$

Figure 1: Forward checking algorithm

$\forall_{1 \leq i \leq m}\ c_i(a) = true.$

A search tree can be used to find a solution to a constraint satisfaction problem (CSP). Each node of the tree corresponds to a partial solution. We get from one node to the next by assigning a value to an uninstantiated variable.

While searching for a solution to a CSP, consistency checking can be used to reduce the size of the search tree and detect dead end paths before all variables have been instantiated. Consistency checking is the process of removing values from the domains of uninstantiated variables which are inconsistent with any solution using the values assigned to the already instantiated variables. That is, values are removed from the domain which violate some constraint. The use of consistency checking techniques is also known as constraint propagation. We will review three such techniques which are described for binary constraints. Binary constraints are constraints which involve only two of the variables. The first two, forward checking and full looking ahead, are described in [HE80]. The third, arc consistency, is described in [HDT92].

Forward checking is the simplest of the three techniques. In forward checking, when a variable x is instantiated with a value v, the domains of all uninstantiated variables are updated by removing values which are inconsistent with $x = v$. The algorithm for forward checking appears in Figure 1.

Full looking ahead is similar to forward checking but involves additional processing to remove more inconsistent values from the domains of the uninstantiated variables. The first step in full looking ahead is to perform the forward checking described above. Then each uninstantiated variable x_i is processed as follows. For each value $v' \in D(x_i)$ we examine the domain of each uninstantiated variable x_j for which there is a constraint between x_i and x_j. If there is no value $v'' \in D(x_j)$ which is consistent with $x_i = v'$ then v' is removed from the domain of x_i. The algorithm for full looking ahead appears in Figure 2.

While full looking ahead removes more inconsistent values than forward checking alone, it does not remove

Let y be the variable just instantiated with the value v.
Let $X = \{x_1, ..., x_n\}$ be the set of uninstantiated variables.
  Perform forward checking.
  For $i = 1$ to n
    For each $v' \in D(x_i)$
      For $j = 1$ to n
        If $i \neq j$ and there is a constraint between x_i and x_j
          then if there is no $v'' \in D(x_j)$ which is consistent with $x_i = v'$
            then $D(x_i) \leftarrow D(x_i) - v'$

Figure 2: Full looking ahead algorithm

all inconsistent values. The third technique, arc consistency, does. In this method, whenever the domain of a variable y is changed, either by instantiation or consistency checking, y is placed in a queue. While the queue is not empty, the following steps are repeated. A variable x is removed from the queue. For each uninstantiated variable x_i such that there is a constraint between x and x_i, each value v' in the domain of x_i is examined to see if there is a value $v'' \in D(x)$ which is consistent with it. If there is no such v'', then v' is removed from the domain of x_i and x_i is added to the queue if it's not already in the queue. The arc consistency algorithm appears in Figure 3.

An example illustrates the different amounts of processing done by the three methods. Consider a graph coloring problem on six vertices u, v, w, x, y, and z with four colors $1, 2, 3$, and 4. The adjacencies are given by the graph in Figure 4. Initially the domains of all six variables are $\{1, 2, 3, 4\}$. Suppose we first assign u the color 1 and v the color 2. Using any of the three algorithms the domains of the uninstantiated variables would be reduced to: $D(w) = D(z) = \{3, 4\}, D(x) = D(y) = \{2, 3, 4\}$.

Now suppose that we assign w the color 3. Using forward checking, 3 will be removed from the domains of x, y, and z. This gives domains of $D(x) = D(y) = \{2, 4\}$ and $D(z) = \{4\}$. Full looking ahead performs forward checking getting these domains. It then processes the variable x checking to see if there are any values in the domain of x for which there are no consistent values in the domain of y since there is an inequality constraint between x and y. No inconsistent values are found in the domain of x. Next the domain of y is checked to see if there are any values which are inconsistent with the domain of z and 4 is removed from the domain of y. Full looking ahead gives these domains: $D(x) = \{2, 4\}, D(y) = \{2\}$, and

Let y be the variable just instantiated.
Let $X = \{x_1, ..., x_n\}$ be the set of uninstantiated variables.
Place y in a queue.
While the queue is not empty
Remove the first variable,x, from the queue.
For $i = 1$ to n
If $x_i \neq x$ and there is a constraint between x_i and x
then for each $v' \in D(x_i)$
if there is no $v'' \in D(x)$ which is consistent with $x_i = v$
then $D(x_i) \leftarrow D(x_i) - v'$ and
place x_i in the queue if its not aready there.

Figure 3: Arc Consistency Algorithm

$D(z) = \{4\}$. If arc consistency is used the domain of x would again be compared against the domain of y and 2 would be removed from x's domain. So with arc consistency the domains are: $D(x) = \{4\}, D(y) = \{2\}$, and $D(z) = \{4\}$.

These constraint propagation techniques can be applied to scheduling problems. When using a search tree for a scheduling problem the nodes represent partial solutions which are obtained either by assigning start times to tasks or by ordering two tasks which require the same resource. Maintaining consistency generally entails reducing the intervals which represent the possible start times of the tasks.

This search method finds a single solution to a scheduling problem. If an optimal solution which minimizes the makespan is desired, this method can be used to find an initial solution. Then the search can be iterated with a new constraint stating that the makespan must be less than the best found so far.

Maintaining additional information can increase the efficiency of scheduling algorithms. Generally time windows are kept for each task which represent the possible times when execution of the task can begin. In [CL94, CL95, CL96] time intervals are kept for sets of tasks which use the same resource. This allows for greater propagation and thus earlier detection of inconsistencies.

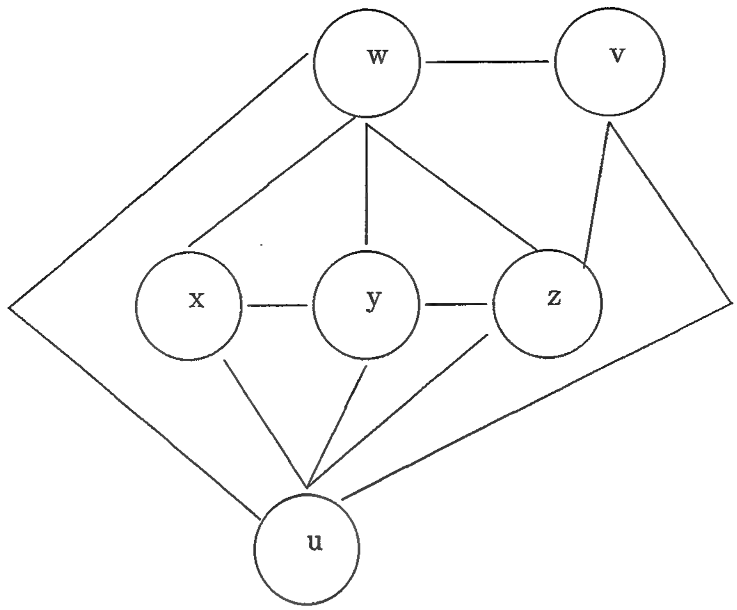


Figure 4: A Graph Coloring Problem

2.2 Tools

Recently tools for constraint reasoning have been developed. One such tool is ILOG SOLVER [2]. SOLVER is a C++ library of constraint reasoning algorithms. The arc-consistency technique described in [HDT92] is used by the SOLVER constraint propagation algorithm. SOLVER also contains built in search mechanisms which implement the tree search technique previously described.

ILOG also offers another C++ library SCHEDULE which can be used in conjunction with SOLVER for scheduling problems. SCHEDULE has predefined classes for tasks and various types of resources. There are also predefined constraint mechanisms for precedence and resource requirements. These two products have been used in implementations for solving the job shop scheduling problem [BPN95a, BPN95b].

3 Benchmarks

Evaluating algorithms empirically can be done by testing them on known benchmarks. There is an extensive collection of benchmarks for the standard job shop scheduling problem. Table 1 shows the optimal value for some of those benchmarks as reported in [CL95]. In cases where the optimal value is not yet known, upper and lower bounds are given.

Aside from the JSSP benchmarks there is a dearth of available benchmarks for testing scheduling algorithms. There are however numerous benchmarks available for linear programming algorithms. Linear programming problems, like scheduling problems, are a subclass of constraint satisfaction problems. As such,

[2] ILOG SOLVER and ILOG SCHEDULE are products of Ilog,Inc., 2005 Landings Dr., Mountain View, CA 94043

Name	size	optimal value	lower bound	upper bound	Name	size	optimal value	lower bound	upper bound
la01	10 x 5	666			la32	30 x 10	1850		
la02	10 x 5	655			la33	30 x 10	1719		
la03	10 x 5	597			la34	30 x 10	1721		
la04	10 x 5	590			la35	30 x 10	1888		
la05	10 x 5	593			la36	15 x 15	1268		
la06	15 x 5	926			la37	15 x 15	1397		
la07	15 x 5	890			la38	15 x 15	1196		
la08	15 x 5	863			la39	15 x 15	1233		
la09	15 x 5	951			la40	15 x 15	1222		
la10	15 x 5	958			orb1	10 x 10	1059		
la11	20 x 5	1222			orb2	10 x 10	888		
la12	20 x 5	1039			orb3	10 x 10	998		
la13	20 x 5	1150			orb4	10 x 10	1005		
la14	20 x 5	1292			orb5	10 x 10	887		
la15	20 x 5	1207			orb6	10 x 10	1010		
la16	10 x 10	945			orb7	10 x 10	397		
la17	10 x 10	784			orb8	10 x 10	899		
la18	10 x 10	848			orb9	10 x 10	934		
la19	10 x 10	842			orb10	10 x 10	944		
la20	10 x 10	902			mt06	6 x 6	55		
la21	15 x 10	1046			mt10	10 x 10	930		
la22	15 x 10	927			mt20	20 x 5	1165		
la23	15 x 10	1032			abz5	10 x 10	1234		
la24	15 x 10	935			abz6	10 x 10	943		
la25	15 x 10	977			abz7	20 x 15		655	665
la26	20 x 10	1218			abz8	20 x 15		638	670
la27	20 x 10	1235			abz9	20 x 15		656	668
la28	20 x 10	1216			yam1	20 x 20		826	888
la29	20 x 10		1130	1157	yam2	20 x 20		861	912
la30	20 x 10	1355			yam3	20 x 20		827	898
la31	30 x 10	1784			yam4	20 x 20		918	977

Table 1: Optimal Values and Bounds for JSSP Benchmarks

Field	1	2	3	4	5	6
Columns	2-3	5-12	15-22	25-36	40-47	50-61
Content	type	name	name	value	name	value

Table 2: MPS data fields format

similar techniques can be used to solve linear programming problems. We will examine in depth one such set of linear programming problems, the Kennington problems introduced in [CHK+90], which even resemble scheduling problems in their scope.

3.1 Linear Programming

The linear programming problem in standard form is

$$\begin{array}{ll} \textit{minimize} & c_1 x_1 + \cdots + c_n x_n \\ \textit{subject to} & a_{1,1} x_1 + \cdots + a_{1,n} x_n = b_1 \\ & \vdots \\ & a_{m,1} x_1 + \cdots + a_{m,n} x_n = b_n \\ & x_i \geq 0 \textit{ for } i \in \{1, \ldots, n\} \end{array}$$

This is often written as

$$\begin{array}{ll} \textit{minimize} & c^T x \\ \textit{subject to} & Ax = b, x \geq 0 \end{array}$$

where c is the coefficient vector for the objective function, x is the vector of variables, A is a matrix containing the coefficients of the variables in the constraints, and b is the vector containing the values for the right-hand-sides of the constraints. The rows of the matrix A correspond to the constraints and the columns to the variables. Thus an element $a_{i,j}$ of A gives the coefficient for variable j in constraint i.

Linear programs specified in forms other than that above may be converted to standard form by various manipulations. For example, slack variables may be added to constraints to transform inequalities into the required equalities. A maximization problem may be solved by formulating the dual minimization problem.

3.2 The MPS format

The Kennington problems are stored in datafiles in MPS format. Mathematical Programming System (MPS) is the industry standard for storing data for linear programming problems. Problems are formulated in a modified standard form which allows inequality constraints. MPS doesn't have a default interpretation of problems as either maximizations or minimizations. The Kennington datasets, as described in Section 3.3, appear to assume minimization as default.

The data in a MPS file is divided into sections. There are five required sections: NAME, ROWS, COLUMNS, RHS, and ENDATA. There are also two optional sections RANGES and BOUNDS. These section keywords must be capitalized and appear in the first column.

The data in the sections is column-sensitive and divided into six fields as shown in Table 2. The meaning of the data in a field depends upon which section it is in. This is best illustrated by an example.

```
NAME          EXAMPLE
ROWS
 L  CON1
 G  CON2
 G  CON3
 N  COST
COLUMNS
    X         CON1        3.          CON2        2.
    X         CON3        1.          COST        2.
    Y         CON1        2.          CON3        5.
    Y         COST        4.
    Z         CON1        2.          CON2        2.
    Z         CON3        3.
RHS
    RHS       CON1        9.          CON2        3.
    RHS       CON3        12.
RANGES
    RNG       CON2        3.
BOUNDS
 UP BND       Z           3.
ENDATA
```

The first section NAME contains only one piece of information, the name of the problem which is contained in the third field. So the name of the problem above is EXAMPLE.

The ROWS section is used to specify the names and types of the constraints which make up the rows of the coefficient matrix. For each line in this section the first field contains an indicator of the type of constraint for the row named in the second field. There are four indicators: N for the objective function and

unconstrained rows, L for $\leq$ inequalities, G for $\geq$ inequalities, and E for equalities. The first N type row is considered to be the objective function. So in the example above COST is the objective function, CON1 is $\leq$ constraint, CON2 and CON3 are $\geq$ constraints. There are no equality constraints in the example.

The section COLUMNS gives the values for each nonzero element of the coefficient matrix. That is, it gives the coefficients for the variables in each constraint. In this section field 1 is not used. For each line of data in this section field 2 contains the name of a column (variable). Fields 3 and 4 contain the name of a row and the value of the coefficient for the column variable in that row. Fields 5 and 6 may contain another row,value pair for the column. In EXAMPLE the variable X has a coefficient of 3 in CON1.

The RHS section specifies the values of the right-hand-side of the constraints. The format for the RHS section is the same as for the COLUMNS section. Field 1 is not used. In field 2 is the name of the RHS vector. In fields 3 and 4 are the name of a row and value of the RHS vector for that row. Fields 5 and 6 may contain another row, value pair. In the example the right-hand-side for CON1 is 9. More than one RHS vector may be specified. By default the first such vector is considered to be the right-hand-side vector for the problem. If a right-hand-side is not specified for a row, then the right-hand-side value is assumed to be 0.

It is possible for constraints to have both a lower and upper bound. Since only one of these may be given by the indicator in ROWS and the information in RHS, the RANGES section allows the specification of the other bound. Field 1 is not used in this section. In field 2 a name is given to the column of ranges. Fields 3 and 4 contain the name of a row and range value. Fields 5 and 6 may contain another row, range value pair. The range value,r, is calculated by subtracting the lower bound from the upper bound. Thus for a $\geq$ inequality we can add $|r|$ to the RHS value for the inequality (which is it's lower bound) to obtain the upper bound for the constraint. Similarly for a $\leq$ inequality we can subtract $|r|$ from the RHS value for the inequality to obtain the lower bound for the constraint. For an equality constraint, if r is negative (respectively positive) it is added to the RHS value to obtain the lower (upper) bound for the constraint with the RHS value taken to be the upper (lower) bound for the resulting inequality. In the example CON2 has a range of 3. Since CON2 is a $\geq$ inequality with a RHS of 3, the constraint has a lower bound of 3 and an upper bound of 6. As with the RHS, more than one RANGE may be specified.

The BOUNDS section allows specification of upper and lower bounds for some or all of the columns (variables). Field 1 indicates what type of bound is being specified for the variable. UP denotes upper bound, LO lower bound, FX a fixed variable, FR a free variable, MI that the lower bound is $-\infty$, and PL that the upper bound is $+\infty$. In field 2 a name is given to the row of bounds. Fields 3 and 4 contain the column name and bound value. By default columns have a lower bound of 0 and an upper bound of $+\infty$. In

Name	rows	columns	nonzeros	bounds	mpc	MPS	optimal value
CRE-A	3517	4067	19054	0	152726	659682	2.3595407e+07
CRE-B	9649	72447	328542	0	2119719	10478735	2.3129640e+07
CRE-C	3069	3678	16922	0	135315	587817	2.5275116e+07
CRE-D	8927	69980	312626	0	2022105	9964196	2.4454970e+07
KEN-07	2427	3602	11981	7204	150525	718748	-6.7952044e+08
KEN-11	14695	21349	70354	42698	928171	4167698	-6.9723823e+09
KEN-13	28633	42659	139834	85318	1836457	8254122	-1.0257395e+10
KEN-18	105128	154699	512719	309398	7138893	29855000	-5.2217025e+10
OSA-07	1119	23949	167643	0	1059475	5388666	5.3572252e+05
OSA-14	2338	52460	367220	0	2359656	11800249	1.1064628e+06
OSA-30	4351	100024	700160	0	4470876	22495351	2.1421399e+06
OSA-60	10281	232966	1630758	0	10377094	52402461	4.0440725e+06
PDS-02	2954	7535	21252	2134	197821	801690	2.8857862e+10
PDS-06	9882	28655	82269	9240	769564	3124272	2.7761038e+10
PDS-10	16559	48763	140063	16148	1313834	5331274	2.6727095e+10
PDS-20	33875	105728	304153	34888	2856653	11550890	2.3821659e+10

Table 3: Statistics for Kennington Problems

EXAMPLE, Z has an upper bound set at 3 in the BOUNDS section. By default, X,Y,and Z all have lower bounds of 0 and X and Y have upper bounds of $+\infty$. Again, more than one BOUNDS row may be entered.

ENDATA signifies the end of the MPS file.

So the example data file can be translated into the following linear programming problem.

$$\begin{array}{lrrl}
\textit{minimize} & & 2X+4Y & \\
\textit{subject to} & & 3X+2Y+2Z & \leq 9 \\
 & 3 \leq & 2X+2Z & \leq 6 \\
 & & X+5Y+3Z & \geq 12 \\
 & & X,Y & \geq 0 \\
 & 0 \leq & Z & \leq 3
\end{array}$$

There are extensions of MPS in which it is possible to specify whether the problem is a minimization or maximization, which right-hand-side to use if more than one is defined, which range column to use, and which bounds row to use. For more information on MPS see [Mur81, Arb83, Gre96, CI].

3.3 The Datasets

The Kennington datasets are comprised of four different types of problems. The brief descriptions here of the problems are taken from [CHK$^+$90]. The PDS problems are multicommodity problems generated by the patient distribution system (PDS) generator. They model the situation in which patients need to be moved from a European theatre to U.S. hospitals. Problem size is a function of the number of days since all the problems have 11 commodities. The CRE problems were generated by the channel routing model for

Name	rows	columns	optimal value	low	high
CRE-A	3517	4067	2.3595407e+07	121.6 sec	285.9 sec
CRE-B	9649	72447	2.3129640e+07	18.9 min	60 min
CRE-C	3069	3678	2.5275116e+07	102.2 sec	206.7 sec
CRE-D	8927	69980	2.4454970e+07	14.6 min	25.8 min
KEN-07	2427	3602	-6.7952044e+08	35.9 sec	49.3 sec
KEN-11	14695	21349	-6.9723823e+09	6.6 min	8.2 min
KEN-13	28633	42659	-1.0257395e+10	20.6 min	27.7 min
KEN-18	105128	154699	-5.2217025e+10	4.4 hrs	6.7 hrs
OSA-07	1119	23949	5.3572252e+05	2.7 min	6.1 min
OSA-14	2338	52460	1.1064628e+06	5.8 min	15.2 min
OSA-30	4351	100024	2.1421399e+06	12.2 min	40.0 min
OSA-60	10281	232966	4.0440725e+06	43.1 min	2 hrs
PDS-02	2954	7535	2.8857862e+10	105.1 sec	148.4 sec
PDS-06	9882	28655	2.7761038e+10	28.2 min	43.7 min
PDS-10	16559	48763	2.6727095e+10	3.3 hrs	4.5 hrs
PDS-20	33875	105728	2.3821659e+10	17.7 hrs	32.9 hrs

Table 4: Execution Times for the Kennington Problems

Europe (CRE) generator. These problems involve finding a set of cargo routes to service European bases. Problem size is a function of the number of bases which need to be serviced, The OSA datasets are instances of the problem where aircraft need to be located for a set of operational support missions. These problems were generated by the operational support airlift model (OSA) generator. Problem size is dependent upon the number of requests for passenger transfer and number of bases. The KEN problems are mulitcommodity network flow problems generated with MNETGN. The Kennington problems appear to assume minimization as the default in their MPS representation. The objective functions for the CRE,KEN, and OSA datasets are named COST, which suggests minimization problems. The KEN problems have negative coefficients for the objective function COST, suggesting that these are actually maximization problems which have been reformulated as minimization problems. [Note that this makes sense because the objective for network flow problems is usually to maximize the flow.]

Statistical information for the Kennington problems, taken from [Netb], appears in Table 3. The number of constraints for a problem is one less than the number of rows because one of the rows is the objective function. So CRE-A has 3516 constraints. The number of variables for a problem is equal to the number of columns. It is unclear what the entries in the column nonzeros represent, but it may refer to the number of nonzero entries in the coefficient matrix. That is, the number of entries in the COLUMN section of the MPS file. The optimal value for the problem is shown in the last column of the table. Note that the optimal value for the KEN problems is negative. If these are actually maximization problems reformulated

as minimizations then the optimal value of the maximization problem is the shown optimal value multiplied by -1, i.e. the same value without the $-$ sign. The range of execution times for solving the problems as reported in [CHK$^+$90] appear in Table 4.

None of the data sets contain a RANGES section so there are no constraints with upper and lower bounds in any of the problems. The CRE and OSA sets do not have any bounds specified. All variables x in these problems are such that $0 \leq x \leq +\infty$ by default. The KEN and PDS sets of data have bounds included for some of the variables. The number of bounds in the BOUNDS section is listed in the bounds column of Table 3. These bounds are all upper bounds. So all variables in these data sets have lower bounds set at 0 and upper bounds which are either specified or $+\infty$ by default.

The Kennington datasets can be obtained from Netlib, a collection of public-domain mathematical software and data which includes problem solvers and data for linear programming. Netlib is accessible via the world wide web at http://www.netlib.org and the Kennington datasets can be found at http://www.netlib.org/lp/data/kennington. The data sets are doubly compressed as explained in [Neta]. They must be copied in binary mod e and then uncompressed with "uncompress". Then they must be uncompressed with "emps", which is available in C and Fortran versions from http://www.netlib.org/lp/data. The number of bytes in the file for a problem after "uncompress" is in column mpc of Table 3 and the number of bytes after "emps" is in column MPS of the table. After performing these steps the datasets will be in the MPS format.

4 Conclusion

We have surveyed several scheduling problems. We also examined the related constraint satisfaction problem and three methods of maintaining consistency while searching for solutions to these problems. Forward checking requires the least amount of time but removes the fewest inconsistent values. Full looking ahead involves more processing time but also removes more inconsistent values that forward checking. Arc consistency removes all inconsistent values but requires the most processing time. We also examined in depth a set of benchmarks for linear programming, the Kennington problems, which could also be used to test scheduling algorithms.

References

[Arb83] Ari Arbel. *Exploring Interior-Point Linear Programming,Appendix A.* MIT Press, 1983.

[Bak74] K.R. Baker. *Introduction to Sequencing and Scheduling.* Wiley & Sons, 1974.

[BPN95a] P. Baptiste, C. Le Pape, and W. Nuijten. Constraint-Based Optimization and Approximation for Job-Shop Scheduling. In *Proceedings of the AAAI-SIGMAN Workshop on Intelligent Manufacturing Systems, IJCAI-95,Montreal,Canada*, 1995.

[BPN95b] P. Baptiste, C. Le Pape, and W. Nuijten. Incorporating Efficient Operations Research Algorithms in Constraint-Based Scheduling. In *Proceedigs of the First International Joint Workshop on Artificial Intelligence and Operations Research, Timberline Lodge, Oregon*, 1995.

[CHK+90] W.J. Carolan, J.E. Hill, J.L. Kennington, S. Niemi, and S.J. Wichmann. An Empirical Evaluation of the KORBX Algorithms for Military Airlift Applications. *Operations Research*, 38(2):240–248, 1990.

[CI] Rice University Computer and Information Technology Institute. "mps_format". Available via anonymous ftp from softlib.cs.rice.edu in /pub/miplib/mps_format. Also available at website ftp://softlib.cs.rice.edu/pub/miplib/mps_format.

[CL94] Y. Caseau and F. Laburthe. Improved CLP Scheduling with task intervals. In P. Van Hentenryck, editor, *Logic Programming, Proceedings of the Eleventh International Conference on Logic Programming*, pages 369–383, Santa Margherita Ligure, Italy, 1994. MIT Press.

[CL95] Y. Caseau and F. Laburthe. Disjunctive Scheduling with Task Intervals. Technical report, LIENS Technical Report 95-25, École Normale Supérieure Paris, France, July 1995.

[CL96] Y. Caseau and F. Laburthe. Improving Branch and Bound for Jobshop Scheduling with Constraint Propagation. In M. Deza, R. Euler, and Y. Manoussakis, editors, *Combinatorics and Computer Science, 8th Franco-Japanese 4th Franco-Chinese Conference*, to appear in LNCS, Brest, France, 1996. Springer Verlag.

[GH96] C.P. Gomes and J. Hsu. ABA: An Assignment Based Algorithm for Resource Allocation. *SIGART Bulletin*, 7(1):2–8, 1996.

[Gre96] John W. Gregory. "Linear Programming FAQ" Usenet sci.answers. Available via anonymous ftp from rtfm.mit.edu in /pub/usenet/sci.answers/linear-programming-faq. Also available at website http://www.skypoint.com/subscribers/ashbury/linear-programming-faq.html, 1996.

[GTT94] C.P. Gomes, A. Tate, and L. Thomas. A Distributed Scheduling Framework. In *Proceedings of the Sixth International Conference on Tools with AI*, 1994.

[HDT92] P. Van Hentenryck, Y. Deville, and C. Teng. A Generic Arc-Consistency Algorithm and Its Specializations. *Artificial Intelligence*, 57:291–321, 1992.

[HE80] R.M. Haralick and G.L. Elliott. Increasing Tree Search Efficiency for Constraint Satisfaction Problems. *Artificial Intelligence*, 14:263–313, 1980.

[Mur81] Bruce A. Murtagh. *Advanced Linear Programming*, chapter 9. McGraw-Hill, 1981.

[NA96] W.P.M. Nuijten and E.H.L. Aarts. A Computational Study of Constraint Satisfaction for Multiple Capacitated Job Shop Scheduling. *European Journal of Opeational Research, to appear*, 1996.

[Neta] Netlib. LP/DATA README. Available at website http://www.netlib.org/lp/data/readme.

[Netb] Netlib. lp/data/kennington/readme. Available at website http://www.netlib.org/lp/data/kennington/readme.

Complexity, Ontology, and the Causal Markov Assumption

Paul B. Losiewicz, Ph.D.

University of Texas at Austin
Department of Philosophy
Waggoner 316 Austin, TX

Final Report for:

Summer Faculty Research Program
Rome Laboratory

Sponsored by:

Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory
Rome, New York

July 1996

ABSTRACT

The question of what constitutes a causal model is significant, as it will lead to a better understanding of the role of causal foundations and heuristics for probabilistic reasoning. It has been argued that in certain cases domain specific considerations can be appealed to in the construction of more efficient causal models that are "non-standard" in the way networks reflect anomalous correlations between nodes. For the most part, the causal assumptions generally invoked [Pearl 1988], [Spirtes, Glymour, Scheines, 1993] do lead to systematic efficiencies based on a reduction in computational requirements for the models they produce. The goal of this paper is to uncover some of the assumptions about causality that undergird current causal models, assumptions which should be kept in mind by those invoking causal relations in the construction of discovery algorithms for causal networks.

Complexity, Ontology, and the Causal Markov Assumption

Paul B. Losiewicz, Ph.D.

1. Introduction

The question of the nature of causality with respect to methods of probabilistic reasoning has been of ongoing interest to the AI community, and the question has recently manifested itself as a criticism of the Causal Markov Assumption (CMA) [Spirtes, Glymour, Scheines, 1993], [Lemmer 1996, pending]. The CMA provides a valuable method for the reduction of complexity via implicit application of notions of causality to models of probability distributions. An inspection of the literature will also show that the Causal Markov Assumption [Friedman and Halpern, 1996], [Elby 1992] is identical to Reichenbach's common cause requirement [Reichenbach, 1956, pp. 188-189]. A better understanding of the foundations of the relationship between causality and probabilistic reasoning is gained by an "in retrospect" reading of Reichenbach [1949] and [1956], with an eye toward understanding implicit causal assumptions as they are invoked in the literature today. The position defended here is that criticism of unconstrained use of the

Assumption gains support[1], and that "non-standard" causal models may be "empirically justified" in specific domains.

According to Friedman and Halpern [1996] one of the inductive guidelines that has been increasingly invoked for modeling a problem causally is the Markov Assumption. Spirtes, Glymour and Scheines [1993] specifically label the application of the MA within the context of directed graph representations the "Causal Markov Assumption". A close reading of Pearl [1988] supports the view that the CMA is a case of importing causal insights into graphic representations. The question to be investigated here is whether the CMA owes its authority to some causal intuition *per se* (whatever that may ultimately be recognized to be), or to an analogical inference over behaviors from a particular domain.

2. The Markov Assumption and Computational Complexity

It can be stated that originally the Markov Assumption (MA) was recognized as little more than a method for reducing the computational complexity of probabilistic reasoning in dynamic systems [Friedman and Halpern, 1996], [Howard, 1971], [Kemeny and Snell, 1960]. The MA can be spelled out as follows within a state system of countably infinite state variables and countably infinite transitions:

1. A state consists of v_n variables contributing to a state s_n. S is the set of possible states of the system.

2. A partially specified dynamic state description consists of a series of state transitions, $s(0)$ ... $s(n)$, also called a trajectory or a run.

---

[1] For background reading on the epistemological foundations for the theory of Probability from Reichenbach's perspective, see [Reichenbach 1949, vii.

3. If the trajectory is not assumed to follow any deterministic transitions, the state of the system at any point n in the series can be specified probabilistically by computing the probability that any one state will be occupied at n given the entire trajectory through point n using formula (1).

$$(1) \qquad P\{s(n+1) = j | s(n) = i,\ s(n-1) = k, \ldots, s(0)=m\}$$

Given a sufficiently large n of past states in the series and number of potential states s_n, the computation can be very complex . Using the Markov Assumption, one can simplify the computation by assuming that only the n-1 state of the trajectory determines the probability of the transition to s(n). That is, only the immediately preceding state of a process determines the present state of a process, which is determined via formula (2)

$$(2) \qquad P\{s(n+1) = j | s(n) = i\}$$

If there are n possible states, the state transition probabilities for any s_x can be tabulated in an n x n matrix, P, which gives the probabilities p_{ij} that a transition from state s_i to state s_j will occur. The probability of a run r, a run of state transitions, is then the product of the individual state transitions $s_{i,j}$,where given s_0,

$$(3) \qquad \Pr(r) = \Pr[s_0 \ldots s_n] = p_{o,1} \times p_{1,2} \times p_{n-1,n}$$

The primary utility of the Markov Assumption is the possibility of recursively applying it to trajectories of arbitrary length, leading to what is termed a Markov Process [Howard, 71], or a multi-step transition probability calculation. Assuming knowledge of an initial state, one can compute the probability of being in a particular end state given n transitions. An n-step transition calculation can be displayed in a probability matrix ,$\Phi(n)$ where:

(4) $\Phi(n) = P^n$

Thus an n-step transition matrix is obtained from a 1-step transition matrix by raising the 1 step transition matrix to the nth power, causing an exponential growth in the complexity of the calculation of the 1-step matrix. However, *the complexity needs only to be increased to the extent of the dependence of the events.* By invoking the Markov Assumption, we assert that dependency for the purposes of calculation of the probability of an event will be limited to *immediate predecessors* of that event, thereby reaping the benefit of reduced complexity in calculating probability distributions within dynamic systems.

The Markov Assumption, a tool for the reduction of computational complexity within dynamic systems, was later employed within non-dynamic probabilistic systems when insight into the analogy between *immediate predecessor* and *causal determinant* was recognized. This eventually led to a generalization of the Markov Assumption to the context of causal diagrams, where the multiplicity of variables contributing to a state were employed to establish the probability of a state, utilizing polytree representations of the causal factors. This is thoroughly explicated by Judea Pearl [1988].

3. Causality: a Primitive or a Hueristic Aid?

Pearl invokes our intuitions about Causality within the context of causal representations in a number of ways, starting with the claim that that Causality is a *primitive* in the language of probability, and ending with describing causality as a *heuristic aid* in the construction of efficient Bayesian Networks. We will examine Causality in both roles.

As a primitive of probability theory, causal intuitions should form part of the foundation for Pearl's Bayesian methodology. As a heuristic aid, causal intuitions should also be capable of being selectively invoked to aid in the construction of useful graphic representations of a particular domain, apart from the underlying graph theoretic considerations. But, as we will see, these intuitions appear to have very different sources.

Pearl states that, as a heuristic aid, Causation is a "language" with which one can talk efficiently *about* certain structures of relevance relations, with the purpose of separating relevancies from irrelevancies, and for the efficient construction of causal nets. But, we must be clear about the sources of the efficiencies here. Causal models as a whole, coming from the perspective that causality is a primitive, *may* provide a capability that could translate into a reduction in computational complexity, but it should necessarily demonstrate an advance in representational capability. If Pearl's heuristic efficiency claim is correct, then application of heuristc causal considerations in the construction of causal graphic representations ought to provide reductions in computational complexity over networks which ignore such assumptions, or treat their networks in a "non-standard" causal manner, in only a domain specific manner. Thus the question of the efficiency contributed by causal considerations of a heuristic nature should be much more specific to particular domain questions. The question of heuristic efficiency has recently been brought up by John Lemmer[2] [Lemmer 96, pending] , insofar as there may be cases where acceptance of a standard causal heuristic device does *not* yield a more efficient model, and may in fact lead to a less efficient model.

Pearl claims that the theory of graphoids shows that a belief network can constitute a *sound and complete* inference mechanism relative to probabilistic dependencies [1988,

---

[2] John Lemmer [Lemmer 96, pending] has proposed a signal based Communicating Causes model that is not fully Markovian, that does not make counterintuitive ontological commitments, and also in certain cases of converging dependency relations, produces a more economical model

14], that is, it identifies, in polynomial time, every conditional independence relationship that logically follows from those used in the construction of the network. He also states that the *essential* requirement for soundness and completeness is that the network be constructed *causally*, "i.e. that we identify the most relevant predecessors of each variable recursively, in some total order, say temporal." As such, soundness and completeness provides causality a "central role" in knowledge organization, particularly in the efficient representation of irrelevancies. It then follows that how causality influences our handling of recursive structures within the probability calculations licensed by particular graphoids should be a direct result of the "primitive" character of causality.

Pearl believes that Causality is one of four primitive relationships, along with Likelihood, Conditioning, and Relevance, of the "probability language" used to articulate the qualitative relationships useful to normal discourse [Pearl 1988, 16].[3] Pearl's position is that causality is indispensable for structuring and specifying probabilistic knowledge, and, because the semantics of causal relations are preserved by the syntax of probabilistic manipulations, the semantics *intrinsically* conforms to commonsense concepts of causation. What are the syntactic structuring contributions of Causal intuitions? Pearl tells us that Causality describes *directional asymmetries* within relevance diagrams, as representations of nontransitive and induced dependencies. Directionality is thus the key to Causality. According to Pearl [1988,19]:

> ...causal directionality conveys the following pattern of dependency: two events do not become relevant to each other by virtue of predicting a common consequence, but they do become relevant when the consequence is actually observed. The opposite is true for two consequences of a common cause.

---

[3] It is also claimed to be ubiquitous in discourse about man's environment. Yet Pearl is inconsistent about the "primitive" status of causal intuitions, and also adopts a reductionist view, proposing a new account of Causation "based solely on the notion of relevance" [Pearl 1988, 18]. This of course conflicts with the aforementioned claim that Causation is a "primitive" of probability language *along with* relevance, and so it is of interest as to what does constitute the essential requirement (s) for soundness and completeness and what does have the central role in knowledge organization, Causation or Relevance or both.

Pearl specifies that the advantage of directed edge Bayesian Networks over undirected Markov networks, is that a Markov network is unable to represent induced and non-transitive dependencies. That is, Bayesian Networks employ a richer representational notation, that of *directed* graphs (digraphs), "where the directions of the arrows permit us to distinguish genuine dependencies from spurious dependencies induced by hypothetical observations." [Pearl 1988,116]. Thus, in defense of the causality as primitive thesis, if a directed edge is the notational method of representing causal relations within relevance diagrams, and Markov networks are assumed to be relevance diagrams, then causality is not based *solely* on the notion of relevance. Causation, which does require reference to relationships of covariance, found in a Markov Net, also needs to includes *asymmetry,* which we find expressed in the directional notation. For a comparison of an undirected or Markov net with a directed Bayesian net refer to figures 1 and 2.

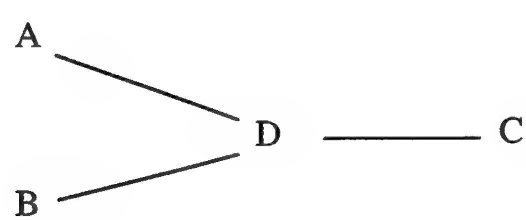


Figure 1. Undirected graph

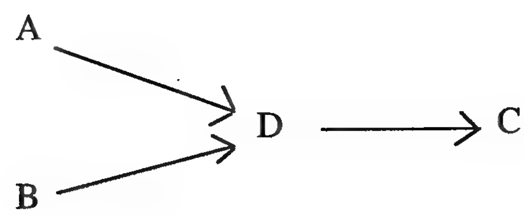


Figure 2. Directed graph

This system wide content is displayed in the "special status" conferred on paths that traverse *converging* arrows. That is, the tails of the convergent diagram are assumed to have no intrinsic dependency relations, but all such dependency relations are *contingent* only on the instantiation of a value for the head. *Induced* relations of dependency, as represented graphically within a configuration called *d-separation*, which configuration exists in both directed and undirected graphs, can only be *adequately* displayed in a directed graph where the depicted dependencies are *asymmetrical* and directionally dependent. Within a Bayesian network, Causality thus has the effect of "modularizing" the experiential knowledge of dependency relations in the domain, and minimizes the number of dependency relations that hypothetically could be considered, by blocking them directionally, and only activating them conditionally as induced dependency relations. Thus when we employ the Markov Assumption within a directed or Bayesian network, the

now *Causal* Markov Assumption does generally provide a savings in computational complexity[4].

It is of historical interest that most of the above was anticipated in Reichenbach [1956]. Reichenbach differs on two points however: First, Reichenbach *clearly identifies* the source of his causal intuitions, which were for him the principles of thermodynamics. Second, a Markov net *does* partially exhibit causal structure insofar as it exhibits the relation of "betweenness", Reichenbach's analogue to Pearl's d-separation, thus allowing an event to *screen off* a second event from a third. This quality is a direct consequence of an *ordered* relation that is represented by a non-directed edge in a graph. This is essential to the screening relation, which places events together in a net but does not establish a serial order. At minimum screening can be represented as a neighborhood relation , but cannot establish a *serial* order for traversing a path through the events [Reichenbach, 1956, 192]. However, that screening *can* occur is a consequence of an ordered structure, a necessary condition of any causal relationship. Furthermore, that events can be connected to multiple events is a requirement of causal structures that are powerful enough to represent the total entropy of a complex system. Thus to this extent a Markov net does display part of the causal structure to be found in Reichenbach's entropy based model of causality.

However, in agreement with Pearl, Reichenbach states that a Markov or non-directed net cannot represent *all* of our causal intuitions, for it is unable to display an *asymmetric* sequence, that is an *irreversible* process. For Reichenbach, it was the irreversibility of the ultimate equiprobable distribution of entropy over the state space that required representation via a *directed* edge. Additionally, it was the possibility of

---

[4] The computational impact of the Markov Assumption on the probability distribution associated with a Bayesian net is formalized by Spirtes, Glymour and Scheines [1993,34] as follows:

representing both convergent and divergent edges that allowed one to model local violations of the second law of thermodynamics by *branching* yet interacting causal subsystems [Reichenbach, 1956, 118]. These complex *causal polytrees* [Pearl, 1988,176] would on the whole demonstrate the overall "statistical isotropy" of the universe [Reichenbach, 1995, 167].

One final structure determining causal intuition derived from Reichenbach has to do with the ultimate ontology of a causal system. Essentially a system is closed, with no possibility of a spontaneous decrease in entropy over the life of the system. All local decreases in entropy *must* be the result of a redistribution of order through an interactive process of the *transfer* of energy from one branch to another. This requires that all probability distributions that display local correlations that exceed chance, i.e. entropic improbabilities, *must* be modeled via a common cause, explicit or latent, so as to be consistent with the commitment to interactive transfer as the *only* source of a decrease in entropy [Reichenbach, 1956,155]. This particular Reichenbachian dictum[5] has also been replicated in Pearl's causal structures as well. Yet Pearl has a tendency to invoke hidden causes for modeling causal anomalies, which mandates an automatic increase in ontology with every probabilistic anomaly. Pearl's propensity for invoking latent causes may reflect his lack of a larger theoretic framework for determining directionality among causal models, and so is tantamount to a heuristic device, explained partially as a result of the expedient of defaulting to a more Baconian approach to data modeling:

Markov Condition: A directed acyclic graph G over **V** and a probability distribution P(**V**) satisfy the Markov condition if and only if for every W in **V**, W is independent of **V** \(**Descendants** (*W*) $\cup$ **Parents** (*W*)) given the **Parents** (*W*)).

[5] "If an improbable coincidence has occurred, there must exist a common cause."[Reichenbach, 1956, 157].

> What is the operational meaning of a genuine causal influence? How do humans distinguish it from spurious correlation? Our instinct is to invoke the notion of control...The test for causal directionality must involve at least one additional variable, say Z, to test if by activating Z we can create variations in Y and none in X, or alternatively, if variations in Z are accompanied by variations in X while Y remains unaltered. [Pearl, 1988,396]

The question is still open as to whether there is any further content, heuristic or otherwise, to be exploited in the concept of causality. It is alleged by Pearl that a Bayesian model of a causal process, as graphically displayed in a Bayesian network, implicitly includes more than the mere incorporation of asymmetrical covariances. If we revert to the *normal discourse* framework usually accepted by Pearl, the network takes on the additional *explanatory* role of displaying real world *reasons* for the putative asymmetrical covariance in the data[6]. Thus we find that there is supposed to be a further role that causality, as reflected in either folk-psychological or expert domain specific causal notions, plays in the construction and evaluation of a specific Bayesian model. What explanatory value contributes to the creation and evaluation of a causal model certainly needs further investigation, and such an investigation will be invaluable in helping understanding the inductive processes that determine the selection of ontologies in modeling a problem.

4. Causal Intuitions as Properly Basic?

As [Howard, 1971,4] states, the Markov Assumption is very strong, yet "No experiment can ever show the ultimate validity of the Markovian Assumption; hence no physical system can ever be classified absolutely as either Markovian or Non-Markovian - the important question is whether the Markov model is useful." But what could constitute a

---

[6] "By contrast the, the arrows in Bayesian networks point from causes to effects, or from conditions to consequences, thus denoting a flow of constraints attributed to the physical world. The reason for this choice is that people often prefer to encode experiential knowledge in causal schemata, and as a consequence, rules expressed in causal form are assessed more reliably." [Pearl 1988,151].

case *against* adoption of the Markov Assumption? Contrary to Howard, Spirtes *et al* [1993, 63ff] admit to the existence of certain experimental results which appear to point to the *falsity* of the assumption within the context of recent quantum mechanical research [Elby, 1992] but add that :

> In our view the apparent failure of the Causal Markov Condition in some quantum mechanical experiments is insufficient reason to abandon it in other contexts. We do not for comparison abandon the use of classical physics when computing orbits simply because classical dynamics is literally false. The Causal Markov Condition is used all the time in laboratory medical and engineering settings, where an unwanted or unexpected statistical dependency is *prima facie* something to be accounted for. If we give up the Condition everywhere, then a statistical dependency between treatment assignment and the value of an outcome variable will never require a causal explanation and the central idea of experimental design will vanish. No weaker principle seems generally plausible; if, for example, we were to say only that the causal parents of Y make Y independent of more remote *causes*, then we would introduce a very odd discontinuity: So long as X has the least influence on Y, X and Y are independent conditional on the parents of X, but as soon as X has no influence on Y whatsoever, X and Y may be statistically dependent conditional on the parents of Y.
>
> The basis of the Causal Markov Condition is, first that it is necessarily true of populations of structurally alike pseudo-indeterminist systems whose exogenous variables are distributed independently, and second, it is supported by almost all of our experience with systems that can be put through repetitive processes and whose fundamental propensities can be tested. Any persuasive case against the condition would have to exhibit macroscopic systems for which it fails and find some powerful reason why we should think the macroscopic natural and social systems for which we wish causal explanations also fail to satisfy the condition. It seems to us that no such case has been made.

Various non-quantum, "macrolevel apparent" violations are examined and disposed of by Spirtes *et al* , yet I feel the issue of the significance of the dispute within quantum physics is dismissed too easily by the above quoted appeal to pragmatic considerations. There are two conjuncts in Spirtes *et al*'s criterion for a serious attack on the CMA; one is that there *be* a macroscopic level phenomena that violates the CMA, and the other is that there should be "some powerful reason *why* we should think the macroscopic natural ... systems fail to satisfy the condition."[italics mine].

On the surface, their dismissal of the results of quantum physics experimentation leads one to assume that in fact such results do not come under the purview of macroscopic systems. However, a simple modification to any of these experiments could be made to have significant macrolevel repercussions, if we tie them into a conventional causal apparatus that *does* have such repercussions. The fact that any system that purportedly violates the Causal Markov Assumption is *experimentally testable* is evidence enough that the system could impinge upon the macroscopic level[7].

The other disjunct arguing for the retention of the CMA as stated by Spirtes *et al* is *why* a system should fail to satisfy the condition. In the case of the test of the quantum Einstein - Podolsky - Rosen correlations examined by Elby [1992], the answer is extremely interesting, and worthy of a digression. Elby argues that any causal analysis of the experimental tests of Bell's inequalities *must* postulate latent causes. The argument rests upon a commitment to a latent cause model that that assumes "temporal and causal locality" for any "causal system". Temporal locality is the requirement that every cause precede its effect. Causal locality is the requirement that the state of a measuring device may not determine the causes of an event well separated from that device [Elby 1992, 22-23]. Within the context of the Copenhagen interpretation of relativity theory, temporal and causal locality as defined by Elby forces one to accept violations of spatial locality, i.e. that signals between objects separated in space must proceed at superluminary (i.e. faster than light) speeds. Elimination of one nonlocality brings out another, given the assumption of the completeness of quantum mechanics. The alternative is to question the completeness of quantum mechanics, the position proposed by [Einstein, Podolsky, Rosen, 1935], and

[7] D. Bohm makes a specific case for this interpretation: "Since the apparatus may be regarded, from a physical point of view, as merely a particular case of the large-scale environment of the system, this suggests a multi-level theory in which the general movement of a system is determined by an equation of motion coupling the large scale to the small scale. A measurement process is then interpreted as a special case of this movement, to which each system is always subject." (Bohm and Bub, 1966).

postulate a latent cause in the system. This alternative is based on the assumption that our traditional "intuitions" of causality, e.g. acceptance of a temporal locality that requires a strong ordering of events within a "causal" sequence, or the non-interactive role of a measuring device or of an observer[8], are properly basic. At one level the debate between Einstein *et al* [1935] and Bohr [1935] can be interpreted as over heuristics, i.e. when it is appropriate to supplement a causal analysis with a latent cause. But we see that the implication of Bohr's position that signals can be propagated at superluminary speeds leads to a questioning of the entire causal syntax, i.e. the directional nature of causal influences. Thus it be that certain domain problems may lead us to accept causal models that are not "standard" at all. It is certainly the case that the heated dispute over the empirical support for violations of Bell's inequalities [Bell, 1966], designed to provide confirmation of Einstein's position, is a case of causal intuitions interfering with observation of a specific domain. However, there is no reason not to investigate what alternative formal tools and theoretic assumptions are necessary to model a "causal" anomaly, for only therein lies the possibility of new "causal intuitions", if Reichenbach is correct in his implicit affirmation of the *theory dependent* nature of causal intuition. Support for this view from within the context of the EPR dispute is provided by Henry Stapp [1994]:

> From a purely logical perspective it seems preferable to accept the uniformity of nature's link between the mathematical and physical worlds rather than to inject , without any logical or empirical reason, our notoriously fallible intuitions about the nature of physical reality.

5. Summary and Conclusion

The development of the theory of probability in the form we recognize today was an outgrowth of the analysis of games of chance given by Cardano, Pascal and Fermat.

---

[8] Henry Stapp points out the implicit *a priorism* of those defending our causal "intuitions" regarding the "factual" nature of the macroscopic: "For there is in this structure no natural breakpoint in the chain of events that leads from an atomic event that initiates the chain to the brain event associated with the resulting observational experience."[Stapp, 1994].

According to Hans Reichenbach, the first recognition of causality as a *statistical* process was in Boltzmann's interpretation of the second law of thermodynamics. The *epistemological implication* for us was that what had been assumed to be strict laws of nature were not different from statistical laws of games of chance, now generalized as *a new type* of physical law. "Thus the concept of probability was related to that of causality, and the concept of the statistical law of nature took its place beside that of the causal law of Nature" [Reichenbach 1949, 6]. However, Reichenbach's choice of the word "beside" is indicative that he saw more in causation than the mathematical tool of statistical correlation.

The statistical approach to description of physical processes was originally assumed to be an "expedient necessitated by human ignorance" caused by our own technical inability, using the methods then available, to track the individual molecules in the gasses of interest to physicists such as Boltzmann. This belief rested on the assumption that there were still strict causal laws operating at the micro-level, though as yet unobserved. However, the development of Quantum Physics soon mandated a reversal of this assumption:

> Whereas the first conception asserts the primacy of the causality concept over the probability concept, the second maintains the primacy of the probability concept over the causality concept; the microcosm seems to be governed only by probability laws, whereas for the macrocosm there result statistical regularities that we take for causal laws and from which the idea of a strict causal determinism has been incorrectly extrapolated. [Reichenbach 1949, 7]

It is clear that for Reichenbach the idea of a *strict* causal determinism of all natural phenomena was a *mistaken extrapolation* from the macrolevel models. However, it is reasonable to maintain that the classical causal model itself co-evolved with the theoretical background and mathematical tools used to manipulate the data associated with macro-phenomena. Yet, even earlier philosophic questions about the nature of the relation

between causation and correlation had led to the belief that our pre-scientific notion of causality was intimately involved with the process of induction over correlation. David Hume's identification of causation with induction over observed correlations reinforced the distinction between induction and deduction by providing an informal test procedure for identifying conclusions about observational matters of fact: i.e. they *lacked certainty*. However, recognition of this property as a necessary characteristic of inductive processes did little to explain the role of causality in inductive inference and *vice versa*. Reichenbach argued that Hume was certainly right to a large extent, i.e. that causation *does* depend upon inductive inference with a resultant uncertainty in those inferences. However, Reichenbach made a much stronger case than did Hume, who merely asserted that causality can be reduced to correlation plus "custom" and psychological conditioning [Losiewicz, 1996]. Reichenbach explicitly stated that causation is *not* merely correlation, because , as Reichenbach noted: "The statement that we already have sufficient knowledge of the determining factors of the problem, still depends on inductive inference" [Reichenbach 1949, 8-9]. Thus, Hume's account was inadequate in its discussion of how a causal process might eventually be modeled by a competent researcher and how its ontology is established.

On the other hand, Reichenbach maintained that causality was an *empirically justified* construct derived from the second law of thermodynamics. Based on Reichenbach's linkage of thermodynamics and causality, it can be argued that causal modeling *is* dependent upon inductive insights into causal structures drawn from a specific domain, in his case Boltzmann's observations of the behavior of gasses. As we saw above, Hans Reichenbach advocated an epistemological position that recognized the existence and necessity of inductive foundations to our formalized descriptions of reality. There is little doubt that we describe our world with the intellectual tools we have at hand, including our "intuitions" of the causal structure of the world. I am in agreement with

Spirtes, Glymour and Scheines that we should justify a counter-intuitive model by powerful reasons before we accept it. But any "powerful" reason will no doubt include an explanatory story (what ever that may be determined to mean) and a formalism that effectively models the previously counterintuitive dynamics of the system. To isolate causal intuitions from the theoretical framework and the mathematical tools developed to support them is a mistake, as Reichenbach knew. Furthermore, failure to appreciate the nature of the inductive process that supports these intuitions is to lock a researcher into a form of *a priorism* which can only impede the development of the formal tools necessary to model an alternative set of intuitions. Contrary to the assumptions of the Rationalist traditions Reichenbach contended with, intuitions do not carry their own justifications with them. Within the context of science causal intuitions must be supported by a complex of formal tools for modeling the dynamics of the systems they envision and a theoretic context within which to discuss and test their implications.

Bibliography

Bell, J. (1966). On the Problem of Hidden Variables in Quantum Mechanics. *Review of Modern Physics*. July 1966 v 38, n 3 pp. 447-452.

Bohm, D. and Bub, J. (1966). A Proposed Solution of the Measurement Problem in Quantum Mechanics by a Hidden Variable Theory. *Physical Review,* July 1966, v. 38, n. 3 pp. 453-469

Bohr, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, October 1935, v. 48 pp. 696-702.

Einstein, A. ,Podolsky, B. and Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, May 1935, v.47 pp. 777-780.

Elby, A. (1992). Should we explain the EPR correlations causally? *Philosophy of Science*, v. 59 pp. 16-25.

Friedman , N. and Halpern, J. (1996) A Qualitative Markov Assumption and Its Implications for Belief Change. *12th Conference on Uncertainty and Artificial Intelligence*. 1996. pp. 264-273.

Howard, R. (1971). *Dynamic Probabilistic Systems*. v.1 New York: J. Wiley & Sons.

Lemmer, J. (1996) *The Causal Markov Assumption, Fact or Artifact?* in review.

Kemeny, J. and Snell, J. (1960). *Finite Markov Chains.* Princeton, NJ: D. Van Nostrand.

Kolmogorov, A. (1950) *Foundations of Probability* NewYork: Chelsea .

Kennedy, J. (1995). On the empirical foundations of the quantum no-signaling proofs. *Philosophy of Science* 62, pp. 543-560.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* San Mateo, CA: Morgan Kauffman.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* 82, 4 pp. 669-710.

Reichenbach, H. (1949). *The Theory of Probability.* Berkeley, CA: University of California Press..

Reichenbach, H. (1956). *The Direction of Time.* M. Reichenbach, (ed.) Berkeley: University of California Press.

Stapp, H. P. (1994). Theoretical model of a purported empirical violation of the predictions of quantum theory . *Physical Review*, July 1994, v.50 pp. 18-24.

A STUDY OF A THREE LEVEL MULTIPLE QUANTUM WELL LASER

Erik F. McCullen
Department of Physics

University of Massachusetts Boston
100 Morrisey Blvd.
Dorchester, MA 02122

Final Report for:
Summer Graduate Student Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

September 1996

A STUDY OF A THREE LEVEL MULTIPLE QUANTUM WELL LASER

Erik F. McCullen
Department of Physics
University of Massachusetts Boston

Abstract

The design of a three level Si/ZnS multiple quantum well is examined. The structure consists of three alternating layers of a Si well and ZnS barrier. A lasing wavelength in the near infrared, λ=1.55μm was desired. Thus an energy difference between upper and lower lasing levels, (E_3 and E_2 respectively), of 800meV was needed. A MATLAB program was used to examine the properties of various well and barrier sizes in order to achieve this difference in energy levels. Once a structure was found that met this criteria other properties of the laser was calculated, including all energy levels, the dipole matrix elements, the scattering rate due to acoustic phonons and polar optical phonons, and the lifetimes of all energy levels.

A STUDY OF A THREE LEVEL MULTIPLE QUANTUM WELL LASER

Erik F. McCullen

Introduction:

In recent years the development of silicon based opto-electronic devices has gained considerable interest. The ability to integrate these opto-electronic devices with conventional silicon electronics (of which manufacturing technology is very well developed) is the main reason for this interest. In this paper I will discuss the Si/ZnS structure and the basic physics of multiple quantum wells. I will then discuss the MATLAB program I have written which, for different well and barrier sizes gives the energy levels and other properties of the laser. Then I will discuss the results of this program applied to one specific configuration which matches the main criteria: that the energy difference between the upper and lower lasing energy levels be approximately 800meV.

Multiple Quantum Wells:

The basic structure of the quantum well is alternating layers of an Si well and a ZnS barrier.

LASER STRUCTURE

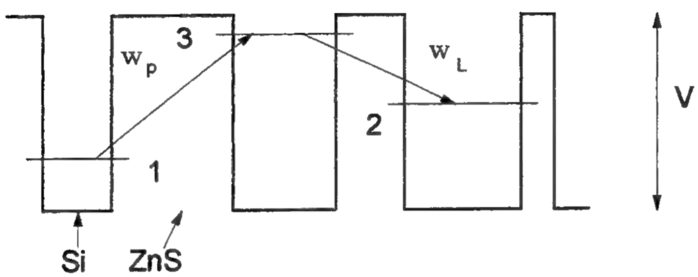


Fig. 1. The basic laser structure; where w_p is the pump frequency, w_L is the lasing frequency and V is the valence band energy offset between Si and Zns.

The wave function in a well n is:

$$\Phi n = An\exp(i\alpha z) + Bn\exp(-i\alpha z)$$

While the wavefunction in a barrier n' is:

$$\Phi n' = An'\exp(\beta z) + Bn'\exp(-\beta z)$$

At the boundary between well n and barrier n' the following continuity equations must hold:

$$\Phi n = \Phi n'$$

and

$$\frac{1}{m*}\frac{d}{dz}\Phi n = \frac{1}{m*'}\frac{d}{dz}\Phi n'$$

and the relationship between the two wavevectors satisfies the following equation:

$$\left(\frac{h}{2\pi}\right)^2 \frac{\alpha^2}{2m*_{Si}} + \left(\frac{h}{2\pi}\right)^2 \frac{\beta^2}{2m*_{ZnS}} = V$$

where the m*'s are the heavy-hole effective masses of Si and Zns, and $h/2\pi$ is hbar. Using the continuity equations the coefficients An, Bn, An', Bn' can be related by a 2x2 matrix T:

$$\begin{pmatrix} An' \\ Bn' \end{pmatrix} = T^n \begin{pmatrix} An \\ Bn \end{pmatrix}$$

Since this is a periodic lattice, at the boundary between one period and the next the following condition holds:

$$\Phi_1 = \Phi_7 \exp(ik_{SL}L)$$

where k_{sl} is the superlattice wavevector and L is the length of one period of the lattice. Imposing these boundary conditions we get the following equation:

$$(T^6 T^5 \ldots T^1 - I)\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = 0$$

The nonzero solution to this equation gives:

$$\det(T^6 T^5 \ldots T^1 - I_{2x2}) = 0$$

The solution to this equation gives the eigenvalues, α_i of the energy of the i'th subband. Then from these α_i's, the β_i's can be calculated along with all the coefficients An, Bn, An', and Bn'. These coefficients then need to be normalized.

To normalize the coefficients A_1 is assumed to be equal to one, and the rest of the coefficients are calculated assuming this. Then the following integration is done:

$$\int(\sum \Phi_i^* \Phi_i) = C^2$$

C must be equal to 1, therefore all the calculated coefficients are divided by C.

Once the normalized wavefunctions are found the lifetimes of all the intersubband levels can be calculated. The two scattering mechanisms used here are the acoustic phonon scattering and polar optical phonon scattering. For acoustic phonons the scattering rate is given by[1]:

$$W_{ij}^a = \left(\frac{\Xi^2 k_B T m^*}{4\pi c_L hbar^3}\right)\int |G_{ij}(q_z)|^2 dq_z$$

Where Ξ is the valence band deformation potential, k_B is Boltzmann's constant, T is the temperature, m* is the valence band heavy-hole effective mass, and c_L is the elastic constant.

For polar optical phonons the scattering rate is given by[2]:

$$W_{ij}^o = \frac{e^2 \varpi_o (n(\varpi_o) + 1/2 \pm 1/2)}{8\pi\varepsilon_p} * \int \left\{ \frac{|G_{ij}(q_z)|^2}{\left[\frac{hbar^4 k^2 q_z^2}{m*^2} + \left(\frac{hbar^2 q_z^2}{2m*} + E_i - E_j \pm hbar\varpi_o\right)^2\right]^{1/2}} \right\} dq_z$$

In this case I consider k=0 therefore the equation reduces to:

Ridley, B.K., Journal of Physics C. **15**, 5899 (1982); P.J. Price, Ann. Phys. (NY) **133**, 217 (1981).

Ridley, B.K. *ibid.*

$$W_{ij}^{o} = \frac{e^2\varpi_o(n(\varpi_o)+1/2\pm 1/2)}{8\pi\varepsilon_p} * \int \left\{ \frac{|G_{ij}(q_z)|^2}{(\frac{hbar^2 q_z^2}{2m*}+E_i-E_j\pm hbar\varpi_o)} \right\} dq_z$$

where ϖ_o is the optical phonon frequency considered a constant, and $n(\varpi_o)$ is the equilibrium number of optical phonons,

$$n(\varpi_o) = \frac{1}{\exp(hbar\varpi_o/k_B T)}$$

and

$$\frac{1}{\varepsilon_p} = \frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_s}$$

where ε_∞ and ε_s are the high-frequency and static permittivity, respectively.

Finally, when these are calculated the lifetime of subband i can be calculated by[3]:

$$\tau_i = \left[\sum_j (W_{ij}^{a} + W_{ij}^{o})\right]^{-1}$$

The MATLAB program:

I used MATLAB to do all the computational work in this paper. Because MATLAB's structure and syntax is more closely like that of normal mathematical symbols, and because of it's graphical abilities it offered an excellent program to examine this laser.

The main program begins by taking in the user's input of the well and barrier sizes. It then begins to scan in energies of 1meV from the bottom of the well to the top of the well. At a fixed energy level the values of α and β are calculated as well as the coefficients An, Bn, An', Bn'. Whenever the matrix

---

Sun, Gang L. Friedman and R.A. Soref, *Appl. Phys. Lett.* **66** (25), 19 June 1995.

equation: $\det(T^6 T^5 T^4 \ldots T^1 - I)$ nears zero within some small tolerance that value of the energy is sent to a MATLAB subroutine 'fzero.m'. This subroutine can calculate the zero of a function down to 15 decimal places.

When the energies reach the top of the barrier, the values of the eigenenergies are sent to a separate file 'eigen.dat'. The values of α and β and the normalized coefficients that correspond to each of the eigenenergies are stored in separate files.

Now that all these values are known, the wavefunctions themselves can be plotted. MATLAB has several simple routines that can plot these functions. Next, the dipole matrix elements $<3|z|2>$, $<3|z|1>$, and $<2|z|1>$ are calculated. These values will be used later in calculating the pumping rate and optical gain. Finally, the main program begins running a smaller program that calculates all the lifetimes of the various energy levels.

A Specific Structure:

The structure shown in figure one is similar to the structure I found that had an energy difference of 800 meV between upper and lower energy states. The only difference being that the middle barrier was the smallest barrier. The sizes of wells being 9, 9.5, and 17.5 Angstroms. The size of the barriers are 20, 12, and 16 Angstroms. A plot of the wavefunctions is given in the end of the paper. The numbers at the top of each graph being the energy level in eV. The x axis being the distance z, in Angstroms. The values used for various constants are:

$$m*_{Si} = .291 m_e$$
$$m*_{ZnS} = 1.76 m_e$$
$$V = 1.5 eV$$

$\Xi_{Si} = 2.6eV$
$\varpi_o = 2\pi * 10.44e^{12}/s$
$c_L = 1.29x10^{11}N/m^2$

This configuration had energy levels of .3019, .6783, and 1.495 eV. The lowest energy level E_1 (.3019eV) was in the first well; the middle energy level E_2 (.6783eV) was in the third well and the highest level E_3 (1.495eV) was in the second well. The difference between upper and lower lasing levels being 816.7meV. The dipole matrix elements were:

$<3|z|2>=4.0483$

$<3|z|1>=8.404e-4$

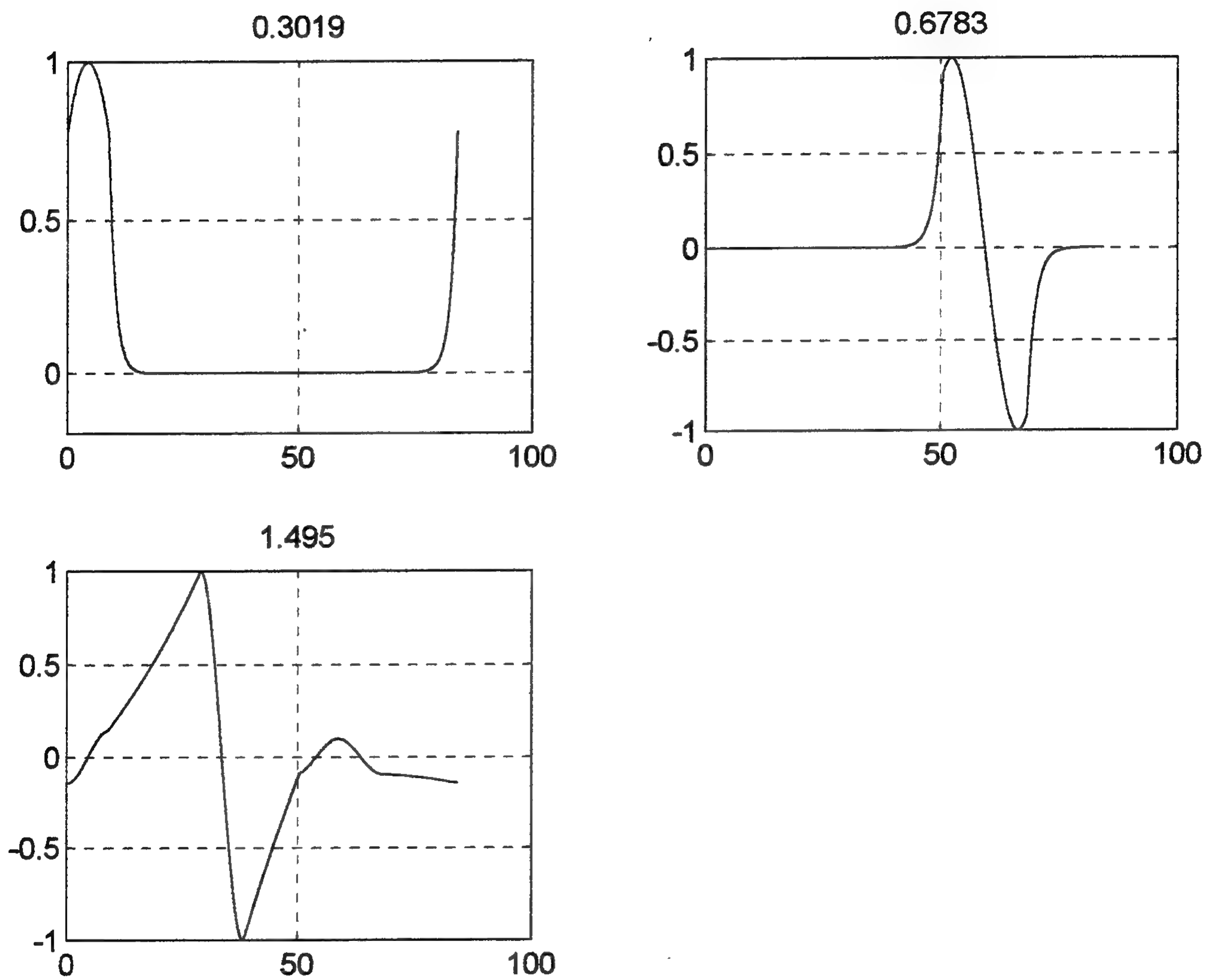
$<2|z|1>=4.3267$

At the time of the writing of this report the lifetimes of the subbands as well as the optical gain of the system had not yet been calculated.

Conclusion:

A program was designed and implemented to examine various configurations of a Si/ZnS multiple quantum well. Using a Si material for an opto-electronic device has been gaining in popularity in recent years. This simple model and program shows that this Si/Zns system may be useful for these devices. This program and the analysis of this system was still continuing at the time of the writing of this paper. Therefore, with a little further work it may show even more benefits for a three-level quantum well laser.

Fig. 2 The wavefunctions for the specific configuration specified in the text. the numbers at the top are the energy levels in eV. The x axis is the distance, z in Angstroms.



References:

Ridley, B.K. Quantum Processes in Semiconductors. Oxford University Press. Oxford; 1993.

Weisbuch, Claude and Borge Vinter, Quantum Semiconductor Structures: Fundamentals and Applications. Academic Press Inc. Boston; 1991.

Gang Sun and Jacob B. Khurgin, "Optically Pumped Four-Level Infrared Laser Based on Intersubband Transitions in Multiple Quantum Wells: Feasibility Study", *IEEE J. of Quantum Electronics*, Vol. **29**. No. **4**. April 1993.

Gang Sun, L. Friedman and R.A. Soref, "Intersubband lasing lifetimes of SiGe/Si and GaAs/AlGaAs multiple quantum well structures". *Appl. Phys Lett.* Vol. **66**. (**25**), 19 June 1995.

Gregory Sun and Lionel Friedman, "Heavy-Hole scattering by confined nonpolar optical phonons in a single Si1-xGex/Si quantum well". *Phys. Rev. B.* Vol. **53**. No. **7**, 15 Feb. 1996.

Experimental Study of Rogowski Profile InP and GaAs Wafers

Jennifer A. Riordan
Graduate Student
Department of Physics

Rensselaer Polytechnic Institute
110 8th Street
Troy, NY 12180

Final Report for:
Graduate Student Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

August 1996

Experimental Study of Rogowski Profile InP and GaAs Wafers

Jennifer A. Riordan
Graduate Student
Department of Physics
Rensselaer Polytechnic Institute

Abstract

This report explores several different processes using Rogowski profile InP and GaAs wafers. There should be no noticeable difference in the photoluminescent (PL) spectrums of the bare wafers from the standard, but with the addition of the passivation layers on these samples, small differences are expected. The PL experiments should give valuable information on the carrier concentration of the intrinsic InP wafer, as well as the passivated ones. With the photoconductive antenna measurements, the results should be comparable to data of typical photoconductive antennas, with an extension to bias voltages exceeding 10 kV. The results of the PL measurements, and their effects on the wafers' success as photoconductive antennas will be discussed, with suggestions for future investigations.

Experimental Study of Rogowski Profile InP and GaAs Wafers

Jennifer A. Riordan

Introduction

Earlier this century, Rogowski described rotationally symmetric, uniform field electrodes through the use of the Maxwell function,

$$Z = \left(\frac{a}{\Pi}\right)\left(W + 1 + e^{W}\right) \qquad (1)$$

with a=G/2V, and G as the desired minimum electrode spacing. This equation represents the equipotential distribution at the end of a parallel plate geometry, extended infinitely along one plane. One finds that the V=0.5Π profile is the greatest amount of curvature that will still provide uniform fields. Although this profile is used for photoconductive semiconductor switch devices much more often as an approximation for two dimensional uniform field contacts, the profile's use in photoconductive antennas has not yet been explored.

Methodology

A. *Wafer Preparation*

Wafer preparation utilizes standard metalization techniques. By cleaning the wafer before continuing, one provides a solid base for the ohmic contacts. Initially, rinse the wafer with isopropanol, then acetone, with a deionized water rinse to complete the process. Fully dry the wafer with compressed nitrogen gas before moving to the photoresist process.

This process is akin to a recipe, which gives the best results when followed. The photoresist itself does not adhere well to the surface of the sample. Applying an adhesive layer of NANO PMGI SF11, Positive Radiation Sensitive Resist 11%CP/THFA from Microlithography Chemical Corp., beforehand

encourages adhesion. By placing the wafer on a Headway Research Photoresist Spinner, adding the PMGI, and then spinning the sample at 5000 rpm for 30 seconds, one attains an even distribution of 1 micron thick adhesive. Place the sample in a 400°C oven for 30 minutes to dry. Upon completion of this phase, the wafer is ready for the photoresistive dye. Apply the AZ4210 Photoresist resin from Hoechst Celanese, a thick dye, in the same manner as the adhesive -- spin at 5000 rpm for 30 seconds to provide a uniform layer of 1 micron, then place in an oven for 20 minutes at 90°C, to densify the resist layer. After completion of this phase, the wafer is ready for the masking phase.

Figure 1 depicts the Rogowski profile, and the mask used in these experiments. Using drawing software, simply model the profile on the computer and print the result on acetate film. The key is to darken the area that will not be exposed to the UV radiation, and lighten that which will indicate the electrodes, that is, a negative-resist. This seems rather simplified, but since the contacts cover such a large area of the wafer, more sophisticated masking techniques are not necessary at this time.

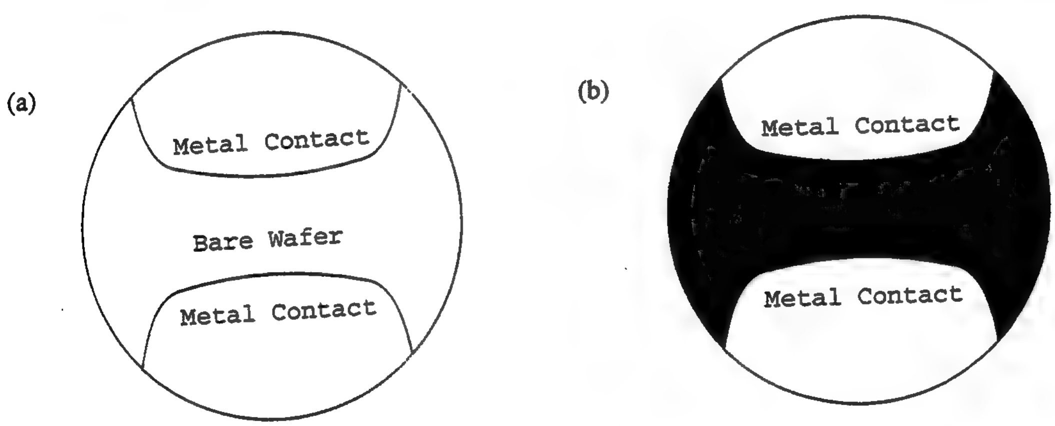


Figure 1 Rogowski Profile. (a) outline form, (b) negative-resist form.

Phase III applies the mask to the wafer. Both are 2 inches in diameter, so alignment is critical. Next, place the sample on the vacuum pedestal in the Solitec Model 3000-IR Mask Alignment and Exposure System with the mask placed on top of the wafer. Turn on the vacuum to stabilize the sample. Solitec's collimated UV arc lamp engages at 15mW/cm^2 for 18 seconds. Upon completion of exposure, place the sample in the Hoechst Celanese AZ400K Developer, diluted to 1:4 Potassium Borate:water for

about one minute, or until the masking effect becomes noticeable and unchanged. Rinse off the developer with deionized water and then dry the sample.

Phase IV is another UV treatment. This time it is a much smaller effect. A Gralab Timer model 171 overhead UV lamp exposes the crystal for 20 minutes. Place in the Microposit SAL-101 developer, an alkali liquid, for 10-15 seconds. The ohmic contacts should attain a slivery finish. Rinse again and dry. An oxide layer develops on the surface from all of these processes, so a quick acid etch becomes necessary. Etch with a 1:1 solution of HCl:deionized water. Within a second or two of immersion, the layer will noticeably emerge from the surface. Deionized water rinse, then dry thoroughly with the compressed nitrogen as before.

Electrode deposition is Phase V, accomplished by placing the wafer(s) in a Comptech Electron-Beam Evaporation System and pumping the chamber down to 10^{-6} Torr. This takes about 30 minutes. The ohmic contacts will be a Ni:Ge:Au:Ni:Au submicron layer, composed of thicknesses of approximately 20:10:20:20:40 nm. In this system, the individual layer thickness must account for the different densities of the materials. As such, recalibration becomes necessary before each layer deposition.

After electrode deposition, which covers the whole wafer, the profile is not distinct. By placing the wafer in an acetone bath, the electrode contacts remain in place, and the rest becomes bare wafer. Remove all the excess metal, then rinse and dry. Place the sample in another developing solution of heated AZ300T from Hoechst Celanese. The remaining photoresist lifts off almost immediately to reveal the final topology of the wafer, then rinse once more in deionized water and dry.

Phase VI is the annealing process, which combines the layers into a single electrode material. Next place the wafer into the 370°C Heatpulse 210 oven, anneal at this temperature for 120 seconds, and then cool slowly to about 150°C. It is only after this exercise that the ohmic contacts are complete.

Phase VII is the passivation[1]. In the present experiments, the benefits of this oxide layer include[8], a surface insulation property that separates the device and metal interconnection, protection of

the diffused junctions from contamination by impurities, and also with an etching of this layer, a contact window forms to allow selective diffusion into the wafer.[1] The samples will be passivated as follows: One will not be passivated, two others will be passivated with an oxide layer, and the last two will be passivated to enhance the surface breakdown voltage.

Phase VIII is the contact window construction. After placing another layer of AZ4210 photoresist on the wafer and heating for 20 minutes at 90°C, expose the wafer to the Solitec UV arc lamp, just as mentioned above. Dip the wafer in the 1:4 developer, as above, then place back in the oven for 1 - 1.5 hours. The masking used in this case must cover most of the contact profile also, allowing only a small window in the ohmic contact region open. Using a buffered hydrofluoric acid solution etch next will remove only that part of the oxide layer not covered by the photoresist after the UV exposure. After rinsing and drying, place the wafer in acetone, which removes the rest of the photoresist, leaving the contact windows in plain view.

B. *Photoluminescence*

Described in the literature[2,3,9,10], photoluminescence is an energy transfer by the absorption of photons. These photons are directly absorbed by the impurity ions in the crystal. Energy level excitation, due to crystal lattice vibrations about the ions' equilibrium, shifts the emission spectrum peak toward the longer wavelengths, known as Stokes' shift. Spectrum breadth is due to a band of absorption or emission wavelengths, not that of a single wavelength. These bands arise from the introduction of the impurity. The conduction and valence bands now have acceptor and donor bands with their own emission activity, thereby increasing the wavelength range. One will also see an additional broadening of the spectrum with an increase in the temperature. Figure 2 shows the experimental schematic. A North-Coast Scientific Corp. Ge-detector, model # EO 817L -250V, serves as the detector in the experiment. There is a high-pass filter before the spectrometer to block out the fundamental 514 nm laser beam.

[1] No specific details on this process are available, as the author did not contribute in this phase.

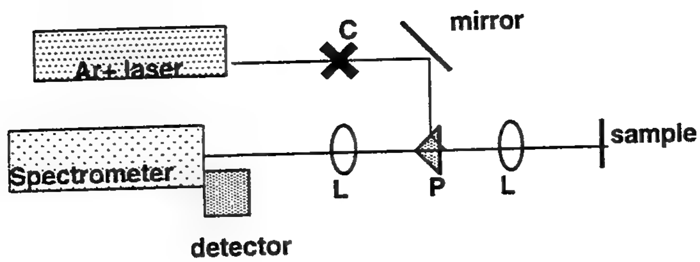


Figure 2 Photoluminescence experiment. L's, lenses; P, prism; C, chopper.

C. *Photoconductivity Measurements*

The basis for the electromagnetic radiation emanating from a semiconductor is the current-surge model[4-7]. An applied electric field accelerates the optically excited carriers, generating an electromagnetic current. This is a surface effect, once attributed to a second-order nonlinear-optical process. This model also predicts that the applied bias field is proportional to the radiated electric field [5,6] with,

$$E_{r,in,max} = -E_b \frac{\sigma_{s,max}\eta_o}{\sigma_{s,max}\eta_o + (1+\sqrt{\varepsilon})} \qquad (2)$$

where $\sigma_{s,max}$ is the peak surface conductivity, expressed as,

$$\sigma_{s,max} = \frac{e(1-R)\mu_{tr}F_{opt}}{h\omega} \qquad (3)$$

where μ_{tr} is the transient carrier mobility at the maximum surface conductivity, R is the optical reflectivity, and F_{opt} is the optical excitation fluence.

Figure 3 shows the experimental set-up of these measurements. The YLF laser, at 527 nm, emits 100 picosecond pulses of 30 mJ pulse energies, chosen by a 400 Hz repetition rate Pockels cell. By inserting the 50 ns delay line, a temporal resolution of 40 GHz occurs, and the obstacle of bandpass limiting is avoided. The weaker split beam triggers the TEK11802 sampling scope, while the loop probe detects the

emitted signal from the sample.

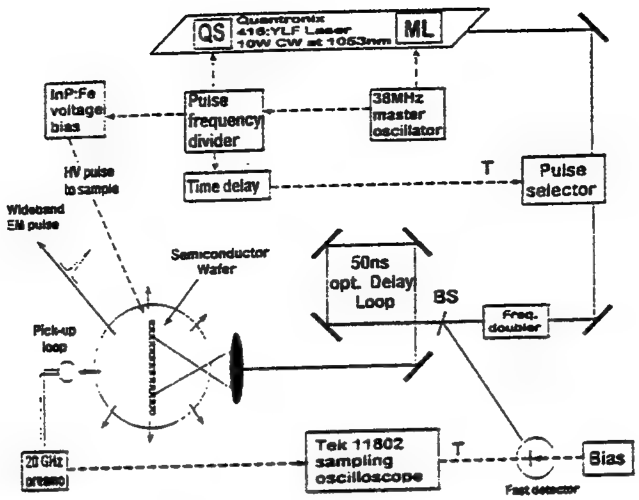


Figure 3 Schematic diagram for the experimental setup[6], used with permission of the author. EM, electromagnetic: T's, triggering: BS, beam splitter.

Results

The InP:Fe wafers were prepared in-house by the Kyropoulos method, while the S-I GaAs wafers were processed by ARL. The GaAs contacts are Ni:Ge:Au:Pd:Au and passivated with SiN, with the processing of the InP wafers outlined above. The spacing of the electrodes in the Rogowski profile is approximately 1 - 1.5 cm.

Throughout the wafer preparation process, several wafers broke at various stages. Therefore, the final number of wafers available dictated the final processing. Three were passivated with the oxide, and none were passivated to enhance the surface breakdown voltage. Also, the non-passivated-wafer broke during an experiment, with no data available.

Photoluminescence (PL) at room temperature is highly dependent on the doping concentration of the substrate. In the case of the InP:Fe, there is possibly an increase in the probability of non-radiative recombinations of photo-generated electron-hole pairs [9], due to an increase in the concentration of deep levels. This implies a decrease in the PL intensity with an increase of the Fe concentration. With this knowledge, and no PL spectrum from either the bare InP:Fe wafer or the passivated sample, there must be

a large concentration of non-radiative recombinations of electron-hole pairs.

From this, one can also determine that as the concentration of deep levels increases, so does the resistivity of the wafer [9] by,

$$\rho \propto \frac{1}{\mu}\left(\frac{N_{Fe}}{(N_D - N_A)} - 1\right) \qquad (4)$$

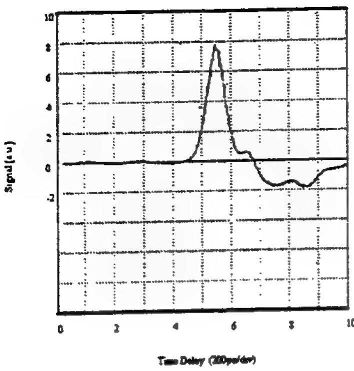
with N_{Fe} as the concentration of deep levels, and μ as the mobility.

The SI-GaAs possesses similar problems, with no PL profile. This contradicts past experiments [10], but the problem may lie in the individual wafers, passivated and not passivated[2]. It is odd, though, not to find any PL activity, as GaAs has a large surface recombination velocity compared to InP. In these experiments, there has been no investigation of the EL2 deep level of GaAs.

For the photoconductivity investigation, the results are more promising. Near-field signals obtained from both the passivated S-I GaAs and the passivated InP:Fe show that both are possible photoconductive antennas. Figure 4 shows the profile of the GaAs wafer, with a 2 kV bias. The pulse width is approximately 150 ps. There is a resistance greater than 1 MΩ across the wafer. Therefore, higher applied bias voltages are possible without damage to the samples.

The InP wafer showed similar results. The measured resistances of the wafers ranged from 0.5 MΩ to about 150 kΩ, with a signal from the 0.5 MΩ sample shown in Figure 5, with a pulse width of 175 ps. In this case, a 4 kV bias applied to the electrodes gave a much smaller signal than the GaAs crystal, with an almost imperceptible signal from the 150 kΩ wafer. Applying much higher biases to the wafers will most likely damage them, as the already high applied bias gives only a small signal. Also, if the incident beam overlaps an electrode by some small amount, the signal shape changes dramatically.

[2] The carrier concentration was later found to be on the order of 10^7 for these wafers.



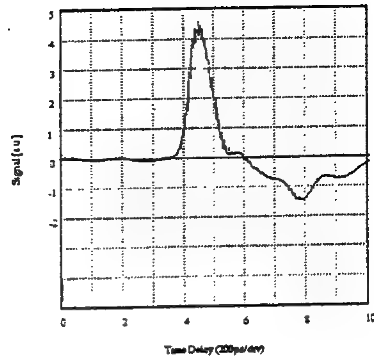


Figure 4 S-I GaAs Signal Pulse

Figure 5 InP:Fe Signal Pulse

In both the GaAs and InP, the profiles show that the applied bias field is proportional to the radiated electric field, as predicted by equation 2.

Conclusions

There is more investigation to be done with the Rogowski profile wafers. By decreasing the carrier concentrations, thereby increasing the surface resistivities of the InP wafers, one could attain better photoluminescent data and photoconductive antennas. By employing the PL data on a bare wafer, the experimenter would be able to enhance the final topology of the wafer through different contacting and passivation.

In the experiments performed here, the InP did not perform as expertly as hoped. . Further experimentation is necessary to determine their success in an ongoing study at Phillips Lab (PL/WSQW), Kirtland AFB in Hydrogen spark gap and photoconductive solid-state (PCSS) technologies. From this cooperation, some further publications and experimentation may blossom in the near future.

<u>References</u>

1. Fink, Donald G., and Christiansen, Donald, **Electronic Engineers' Handbook**, Third Edition, McGraw-Hill Book Co., New York, 1989.

2. Lerner, Rita G., and Trigg, George L., **Encyclopedia of Physics**, Second Edition, VCH Publishers, Inc., New York, 1991.

3. Wilson, J., and Hawkes, J.F.B., **Optoelectronics: An Introduction**, Prentice-Hall International, Inc., London, 1983.

4. Benicewicz, P.K., *et. al.*, **Journal of the Optical Society of America, B**, Vol. 11, No. 12, December 1994, pp. 2533.

5. Darrow, Justin T., *et. al.*, **IEEE Journal of Quantum Electronics**. Vol. 28, No. 6, June 1992, pp. 1607.

6. Liu, D.W., *et. al..*, **IEEE Photonics Technology Letters**, Vol. 8, No. 6, June 1996, pp. 816.

7. Liu, D.W., *et. al..*, **Optics Letters**, Vol. 20, No. 14, July 15, 1995, pp. 15446.

8. Abdalla, Michael D., and Skipper, Michael C., **Photoconductive Semiconductor Switch Experiments**, Prepared for Phillips Lab, Kirtland AFB by Mission Research Corp., Unpublished Report.

9. Erman, M., *et. al.*, **Journal of Crystal Growth**, Vol. 96, 1989, pp. 469.

10. Matsumura, T., *et. al.*, **Journal of Applied Physics**, Vol. 57, 1985, pp. 1182.

AN ATM ADAPTATION LAYER PROTOCOL DESIGNED TO TRANSMIT QUALITY-CRITICAL TCP TRAFFIC OVER DEGRADED COMMUNICATION LINKS

Timothy A. Terrill
Graduate Student
Department of Electrical and Computer Engineering

University at Buffalo
Department of Electrical and Computer Engineering
School of Engineering and Applied Sciences
Bell Hall
Box 602050
Buffalo, New York 14260-2050

Final Report for:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

August 1996

AN ATM ADAPTATION LAYER PROTOCOL DESIGNED TO TRANSMIT QUALITY-CRITICAL TCP TRAFFIC OVER DEGRADED COMMUNICATION LINKS

Timothy A. Terrill
Graduate Student
Electrical and Computer Engineering
University at Buffalo

Abstract

When multimedia applications utilize the TCP/IP suite with ATM over high bit-error-rate (BER) links, very poor performance is the result: TCP connections close themselves because of time-outs. These time-outs are the result of TCP's retransmission waiting time mechanism. The retransmission waiting time timer, for a given segment, is doubled each time it times-out, eventually becoming large enough to cause the connection to close itself. Experiments conducted at Rome Laboratory, involving applications that used TCP to transmit data-critical electronic whiteboard and imagery information over a channel with a BER of 4E-6, concluded that TCP connections consistently closed themselves due to time-outs. Since actual tactical links operate with BERs as low as 1E-3, a new tactical AAL capable of working over high BER links needed to be designed. AALx is a proposed solution to this problem. AALx uses the selective repeat ARQ to provide retransmissions to lost or errored frames. The size of the ARQ frame is adaptable to allow maximum performance to be achieved over the wide variety of communication links.

AN ATM ADAPTATION LAYER PROTOCOL DESIGNED TO TRANSMIT QUALITY-CRITICAL TCP TRAFFIC OVER DEGRADED COMMUNICATION LINKS

Timothy A. Terrill
Graduate Student
Electrical and Computer Engineering
University at Buffalo

Introduction

Multimedia applications will increase the efficiency of military command and control by providing advanced services such as video tele-conferencing and tele-medecine. Some of the benefits of these multimedia applications are: the ability of a commanding officer at headquarters to video tele-conference with his troops in the tactical arena; the ability of a surgeon in a remote hospital to provide medical advice to a medical team located in a helicopter. Heterogeneous computing allows different computer architectures to execute the same applications which can then communicate with each other. Distributed computing is the ability of a network to distribute its resources across the entire network. For example, the base headquarters could be equipped with Sun SPARC workstations using the Solaris operating system, the troops and medical team could be equipped with IBM Laptops using the Linux operating system, and video clips of the target areas or surgical procedures could be located on a video server somewhere else on the network.

Multimedia. The direct translation of the word multimedia, according to Webster's New College Dictionary, is: "using, involving, or encompassing several media". In computing terms it translates into an application that integrates many forms of data into one application; more specifically voice, video, still imagery and raw data. Each of these forms of data require different services from a network. Voice and video require time-critical services. Still imagery and raw-data require quality-critical services.

Time-critical services guarantee delivery of data packets within certain delay and jitter specifications, but do not guarantee error-free delivery. Delay is the total amount of time it takes a data packet to reach the destination once sent onto the network. Jitter is the deviation in the delay experienced by different packets being

transmitted from a single source to the same destination. The Universal Datagram Protocol (UDP) is the protocol that is used to handle time-critical data.

UDP has problems similar to TCP when used in errored environments: UDP packets rarely reach their destination. Experiments conducted at Rome Laboratory concluded that for an application using UDP to transmit time-critical data over a communication link with a BER of 1E-4, the probability of a voice packet being discarded at the AAL level was 25% and the probability of a video packet being discarded at the AAL level was a staggering 97%! AAL0 is the proposed solution to these problems. AAL0, also called the Null AAL, simply passes IP datagrams down to the SAR level performing no computation. The development of the AAL0 protocol is not discussed in this paper. This information was introduced to illustrate the "big picture" of ATM in the errored environment.

Quality-critical services guarantee error-free delivery of data packets with no guarantee on the delay or jitter. Retransmission algorithms and error detection/correction schemes are needed to support these services. The Transmission Control Protocol (TCP) is the protocol that is used to handle quality-critical data.

To efficiently utilize multimedia applications in a heterogeneous, distributed computing environment, powerful networks are necessary. In addition, the network must be able to provide the various types of services required by multimedia applications: time- and quality-critical services. Asynchronous Transfer Mode (ATM) networks provide the solution to all of these requirements. The communication medium primarily used with ATM is fiber. Fiber has very low BERs, on the order of 1E-10.

The military wishes to extend the use of ATM technology into the tactical environment. The tactical environment is composed of wireless communication media, primarily radio and satellite . Radios have extremely high BERs, on the order of 1E-3, and satellites have BERs on the order of 1E-7. Here lies the problem. How can ATM be integrated into the tactical environment when ATM was designed to be used with fiber?

Problem Definition

Most of the current multimedia applications make use of the TCP/IP protocol stack. When these protocols are used on top of ATM, the protocol stack that results

has many problems. Each of the protocols involved in the quality-critical data path will be discussed starting with the ATM layer and working upwards in the protocol stack ending with TCP. Refer to *Figure 1: Protocol Stack of TCP/IP over ATM* to aid in visualization. The figure shows both the time-critical and quality-critical data paths, although the scope of this paper is only the quality-critical data path. The left side of the figure is the quality-critical data path and the right side is the time-critical data path. Notice that there are three levels that do not have separate functionality: the IP, ATM and Physical layers. All data flows through these three levels. Only at the TCP-UDP and AAL3/4-AAL5 levels do the data streams split.

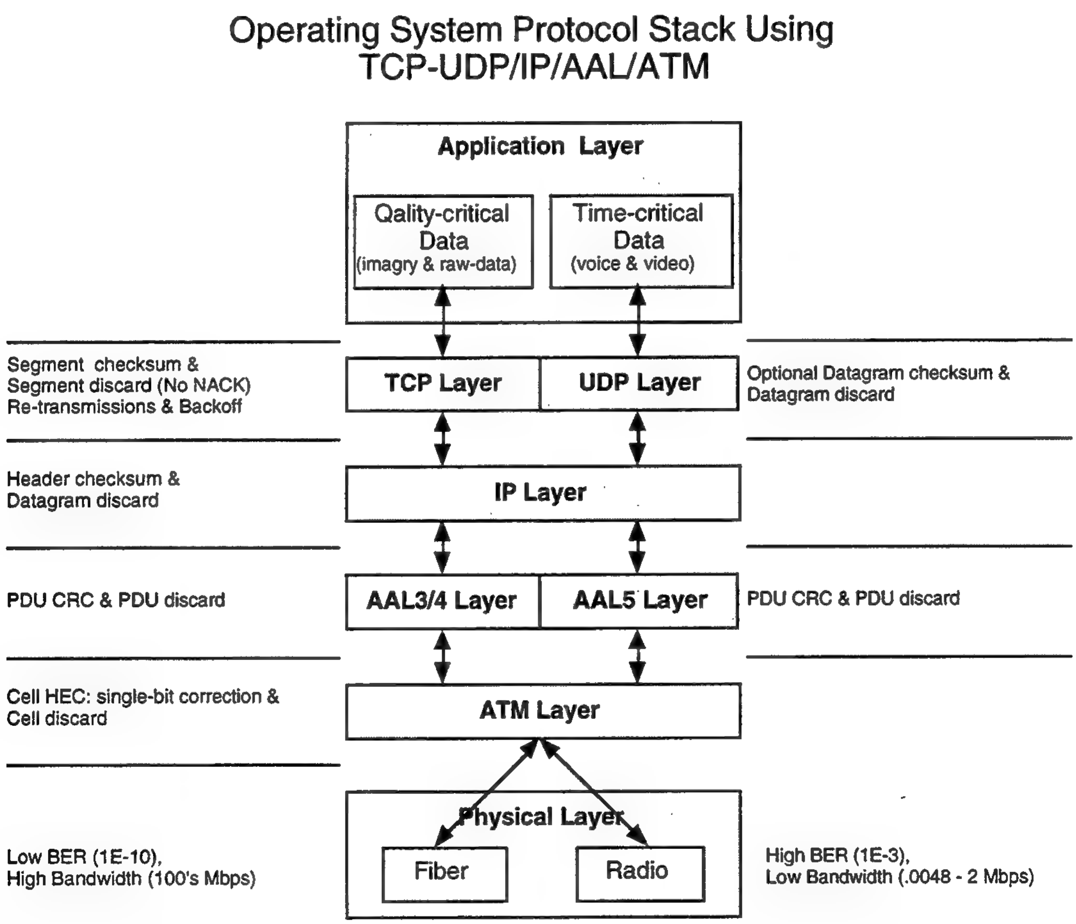


Figure 1: Protocol Stack of TCP/IP over ATM

As discussed in the introduction, ATM was designed to be used with fiber. Since fiber is a very reliable medium, very low BER, ATM was not designed to be a reliable protocol. Single-bit error correction and multiple-bit error detection of the ATM cell header is provided by ATM's header error check (HEC). If there are multiple bit-

errors in the ATM header, the HEC will catch them, but be unable to correct them. The result: the ATM cell will be discarded.

The AAL3/4 layer is made up of two sublayers—the Convergence Sublayer (CS) and the Segmentation and Reassembly (SAR) sublayer—as seen in *Figure 2: The AAL 3/4 Layer.* The CS encapsulates an entire packet from the upper layer, a TCP segment in this case, with header information that allows loss and gain to be detected (the Length field). The CS then passes the resulting packet (called a CS-Protocol Data Unit or CS-PDU[1]) down to the SAR . The SAR does as its name suggests; it segments the CS-PDU into 44-byte chunks, adds its 4-byte header information, and sends the resulting 48-byte SAR-PDU down to the ATM layer. The header contains some fields that are used for error detection: a sequence number (SN), a Length field (LI), and a 10-bit cyclic redundancy check (CRC) that covers the entire SAR-PDU (header, payload and trailer). If, a SAR receives a SAR-PDU and these fields indicate an error, the SAR-PDU is discarded. Once every SAR-PDU corresponding to an entire CS-PDU has been received, indicated by an End Of Message (EOM) SAR-PDU, the SAR informs its local CS. At this point, the CS uses its Length field to detect loss or gain. Loss being the absence of a SAR-PDU that causes the length of the CS-PDU to be shorter than indicated. Gain being the addition of SAR-PDUs that were not part of the original CS-PDU, making the length of the CS-PDU longer than indicated. Gain occurs when header errors pass undetected. If either loss or gain is detected, the entire CS-PDU is discarded.

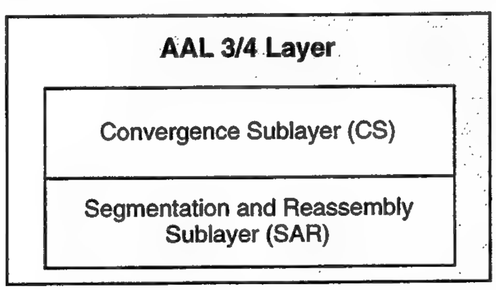


Figure 2: The AAL 3/4 Layer

---

[1] The information passed from one protocol layer to the one below it on the protocol stack has two names. Let's use the CS and SAR layers for an example. The information passed from the CS down to the SAR is called the CS-PDU from the CS perspective. The same information is referred to as the SAR-SDU (Service Data Unit) from the SAR perspective. Both of these naming conventions—PDU & SDU—will be used throughout the paper.

The IP protocol uses a checksum that covers the IP datagram header. If the header checksum detects any errors, the entire IP datagram is discarded. There are no error correction capabilities of the header checksum, error detection capabilities only.

The TCP protocol incorporates a positive acknowledgment, sliding-window, ARQ mechanism designed to handle end-to-end delay, congestion, and lost segments (due to routing errors caused by software bugs or bit-errors in headers, or over full buffers). TCP was not designed to handle link errors; it relied on the Data Link Layer (Ethernet, TokenRing, etc.) for that. In order to support the ARQ mechanism, TCP utilizes many components: a TCP segment checksum, a sliding-window, segment sequence numbers, and retransmission waiting timers (RWTs). For each segment that TCP sends, it calculates a checksum across the entire segment (the unit of transfer of TCP), issues a sequence number (actually a range of sequence numbers, one number per byte being sent) and initiates the countdown of its RWT. The RWTs for all transmitted segments are checked periodically. If one or more have expired (called a time-out), TCP retransmits the segments, increments the values of the RWTs (usually doubles the value up to a maximum value that is set in the operating system kernel), and restarts the RWTs. The sliding window used by TCP is part of the flow control mechanism. Flow control basics are simple; bytes whose sequence numbers fall in the window are allowed to be transmitted. When the window slides, it allows new information to be transmitted. By controlling the size of the window, a host can control the rate at which packets are sent.

A segment that is received containing bit-errors is discarded and no negative acknowledgment (NACK) packet is sent back to the transmitter; this causes a time-out to occur. A NACK is a message that tells the transmitter that the packet was received, but it contained errors, so the packet must be retransmitted.

When a segment is correctly received, a positive acknowledgment (ACK) is sent back to the transmitter which tells the transmitter that the packet was successfully received. Upon reception of an ACK, the window slides ahead and the RWT for that segment is stopped.

Now that all the protocols have been described, lets highlight the problem. Assume a bit error occurs in the ATM payload. ATM will pass the payload up to the SAR. The CRC at the SAR layer will catch the error and discard the SAR-PDU. The CS will detect loss in the CS-PDU and discard the CS-PDU. The transmitting TCP will not receive an ACK for that segment (since it was never received at the other end).

This will cause the RWT to eventually time out. If this were to occur frequently enough, the RWT would eventually become so large that in the process of waiting for the RWT to expire, the TCP connection would time-out (not to be confused with the RWT timing-out), causing the connection to become closed.

A similar result occurs if multiple bit-errors occur in the ATM header. In this case the HEC will be unable to correct them and the ATM cell will be discarded. The SAR layer will notice that a cell is missing and will discard the SAR-PDU. The rest occurs as explained above.

When wireless links are used beneath the ATM layer, large bit-error rates (BERs) on the order of 1E-3 are experienced. When comparing that with the BER for which ATM was designed, fiber, 1E-10, the problem should become apparent. There needs to be some way to present TCP with a reliable segment stream. There are currently provisions for what are known as *Assured Operations* in AAL 3/4. The provisions, described in the CCITT I3.63 - B-ISDN Adaptation Layer (AAL) Specification, say that every assured AAL-PDU payload is guaranteed to be delivered error-free to the destination. There are not, as of yet, any specific implementation details for these assured operations.

Solution

Since AAL3/4 assured operations have not been implemented at the time of this paper, a military AAL, which will provide a reliable segment stream to TCP, must be designed. This AAL, called AALx, will need to implement some sort of automatic repeat request (ARQ) capabilities that enable it to retransmit lost or errored packets. In addition, AALx should be adaptable; it should be able to dynamically determine the best configuration to use for any given link.

Two well-known ARQ protocols were considered for the base development of AALx: the selective repeat ARQ and the go-back-n ARQ.

Both protocols are positive acknowledgment, sliding window protocols. When one end receives an error-free frame, an ACK is sent to the transmitting node containing the requested sequence number (RN) of the next expected frame. Under error-free operation, both protocols operate identically; each use a window of size *n* and that window continually slides forward until the last frame is sent. For both protocols, the window size is determined the same way; the product of the bandwidth and the delay of the link is divided by the average frame size. The window size must

be chosen carefully or else performance hits will be taken; bandwidth will be lost. Also, the size of the sequence number space should be chosen carefully. If it is unnecessarily large, overhead inefficiencies will result. If it is to small, bandwidth will be lost due to the restriction on the window size.

The major difference is in the behavior of the protocols when errors are introduced. The go-back-n protocol cannot receive out of order frames. When a frame is received in error, it discards that frame, all subsequent frames, and transmits the RN of the first errored frame for each newly received frame. The transmitting node continues to transmit frames, sliding its window upon reception of a RN that is greater than the lower edge of the window. When the transmitting node receives a RN that equals the lower bound of the window, it must retransmit the entire window. Thus, go-back-n retransmits an entire window for a single errored frame! Imagine using go-back-n on a satellite link (large delay) where the window size is hundreds or thousands of frames!

Selective repeat is a more efficient protocol in this respect. Selective repeat can receive out of order packets. For each errored frame received, a NACK is sent back containing the RN corresponding to the sequence number (SN) of the errored frame. Upon reception of a NACK, selective repeat retransmits the requested frame. Thus, selective repeat only retransmits a single frame for each errored frame. If the receiver does not receive an ACK within a prescribed time, the transmitter begins retransmitting the entire window. In order to prevent this, the transmitter is allowed to transmit up to 2n frames. The receiver, however, is only allowed to buffer n of them.

To shed a little quantitative light on this subject, results from experiments conducted by the researchers of Harris Corporation, for the Naval Research Laboratory, have been tabulated in *Table 1: Performance Comparison of Selective Repeat ARQ vs. Go-Back-N ARQ over a Geosynchronous Satellite.* The table contains three link specifications (DS-1, DS-3, and OC-3) for each protocol and the link BERs necessary to allow the protocols to achieve efficiencies of 80%. The frame size for this experiment was 30 ATM cells (1590 bytes). For purposes of efficiency calculation, the protocol overhead was that produced by TCP/IP over AAL5. The results show, quite clearly, that the selective repeat protocol performs better over a variety of link bandwidths.

	Selective Repeat			Go-Back-N		
Link Type	DS-1	DS-3	OC-3	DS-1	DS-3	OC-3
BER	10E-5	10E-5	10E-5	5x10E-7	2x10E-8	5x10E-9

Table 1: Performance Comparison of Selective Repeat ARQ vs. Go-Back-N ARQ over a Geosynchronous Satellite

There are many ways to determine if a frame has been errored. Two popular methods, parity checks and cyclic redundancy checks (CRCs), are considered here. For a given header field size, a CRC has much better coverage than the basic parity check. Your basic parity check code can only detect an even number of bit errors; 50% error coverage.

Results

The major components of AALx have been identified and described above, but the low-level details have been left out. Here we discuss both the final design decisions and the low-level details. First we shall cover the low-level details to provide a better understanding of why certain design choices were taken over others. A flow diagram of the operating system protocol stack, with AALx in place of AAL3/4, is shown in *Figure 3: AALx Operating System Protocol Stack.* Each of the layers is identified by a shaded box.

Operating System Protocol Stack Using AALx for Quality-Critical Application Data

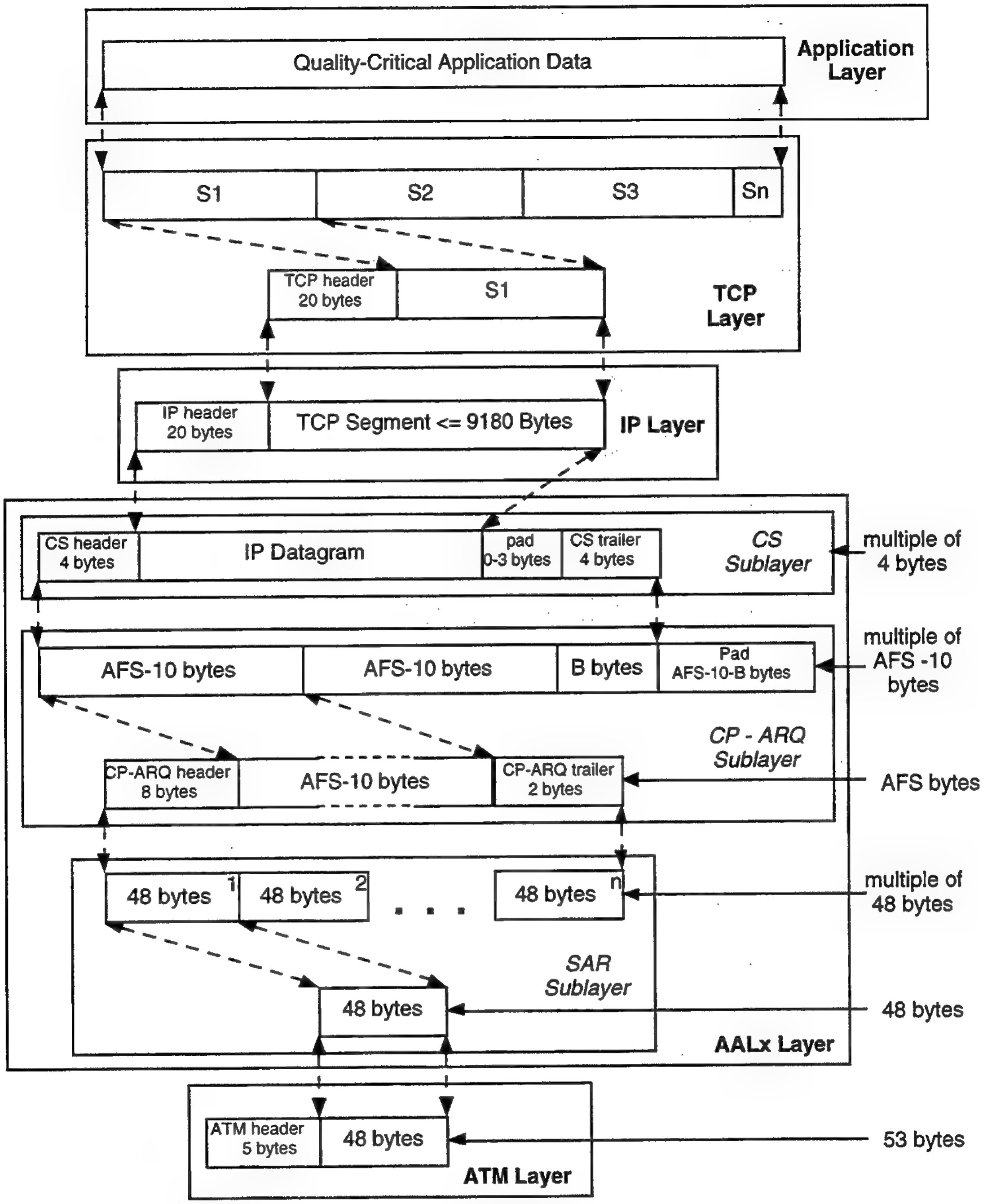


Figure 3: AALx Operating System Protocol Stack

Before actually designing the protocol, it was necessary to analyze the current TCP / IP / AAL3/4 / ATM protocol stack to see where and how the AALx protocol would fit. A quick glance at Figure 3 reveals that there are three sublayers at the AALx level.

It was said earlier that there were only two at the AAL level. Well, for the purposes of what we need to do here, an additional sublayer has been added: the *Common Part Automatic Repeat Request (CP-ARQ)* sublayer. If we had used the existing AAL3/4-CS and added to it the ARQ components mentioned, we would be unnecessarily increasing the complexity of an already complex protocol. For this reason, the CP-ARQ sublayer was designed to perform all the ARQ functions as well as segmentation of the CS-PDU.

Why segmentation? The researchers at Harris Corporation did further experiments on both the Selective Repeat ARQ and the go-back-n ARQ. In these experiments, they varied the size and BER of packets being transmitted across a link. The results showed that for each BER, a peak in the efficiency of the link occurred when a certain packet size was used. This peak occurred at different packet sizes for each BER. For efficiency calculations in their experiment, the overhead of TCP/IP/AAL5/ATM was assumed. Therefore, to obtain the maximum efficiency of a tactical link, it is necessary to have a configurable frame size.

In order to break the CS-PDU into smaller sized chunks, the CP-ARQ layer must do as the TCP layer does; take the information from the upper layer, buffer it, and break it up into smaller pieces. Those pieces are going to be called *ARQ Frames.* In order to make the job of implementing the AALx-SAR simpler, it was decided to make ARQ Frames multiples of 48 bytes. This makes the AALx-SAR identical in operation to the AAL5-SAR; take the SDU, segment it into 48 byte SAR-PDUs, and send each SAR-PDU down to the ATM layer. See *Figure 4: AALx-SAR Frame Structure.*

AALx: SAR - PDU Structure

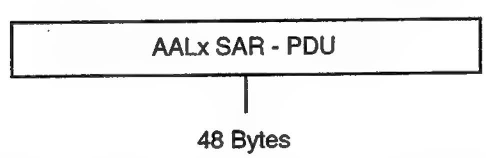


Figure 4: AALx-SAR Frame Structure

We said that the ARQ frame size (denoted AFS) was going to be a multiple of 48 bytes. We also said that the AFS was going to be adaptable. In order to do this, there has to be a mechanism to ensure that a receiving AALx knows which AFS the sending AALx is using. An encoding scheme has been developed that allows this. There will be a field in the CP-ARQ header called the *AFS Factor* that gives the size of

each ARQ frame in 48 byte chunks. That is, the size of each ARQ frame, in bytes, can be determined by:

$$(\text{AFS Factor}) * (48).$$

This encoding scheme saves bits in the header by using a factor instead of transmitting the entire frame size. To determine the size of the AFS Factor field, it was necessary to determine the largest possible AFS. The CS-PDU can have a maximum payload of 65535 bytes and it has a header size of 4 bytes, a trailer size of 4 bytes and since it will be padded on 4 byte boundaries, it will also contain a 1 byte pad. The total, therefore, is 65544 bytes. We need to satisfy

$$(\text{AFS Factor}_{max}) * (48) >= 65544$$

so

$$\text{AFS Factor}_{max} >= 1365.7.$$

To cover 1366, 11 bits are needed. With 11 bits, the coverage is 0 - 2^{11}, which is 0 - 2047. The AFS Factor field will be 11 bits wide. Refer to *Figure 5: AALx-CP-ARQ Frame Structure*, to see the CP-ARQ frame structure. Any values of AFS Factor above 1366 means that an error has occurred. A possible feature could be to check the value and if it is greater than 1366, discard the frame and request a retransmission.

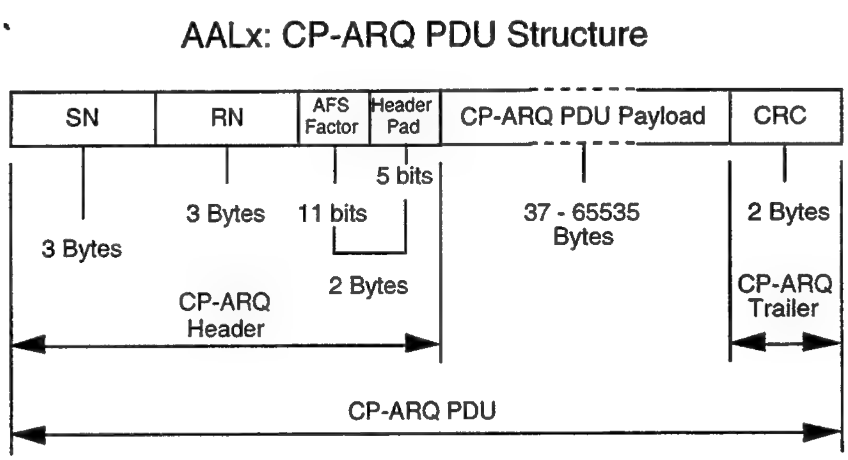


Figure 5: AALx-CP-ARQ Frame Structure

There are a few possible ways for the sending CP-ARQ to determine the AFS that should be used across a given link. There could be a table set up by the system administrator that contains Permanent Virtual Circuits (PVCs) and the specified AFS to be used, as in *Table 2: AFS Factor Lookup Table by PVC Connection.*

PVC	AFS Factor
1	256
4	512
12	32

Table 2: AFS Factor Lookup Table by PVC Connection

The table could also have the AFS decided by the destination host, as in *Table 3: AFS Factor Lookup Table for Host C*.

Host	AFS Factor
A	256
B	512
D	32

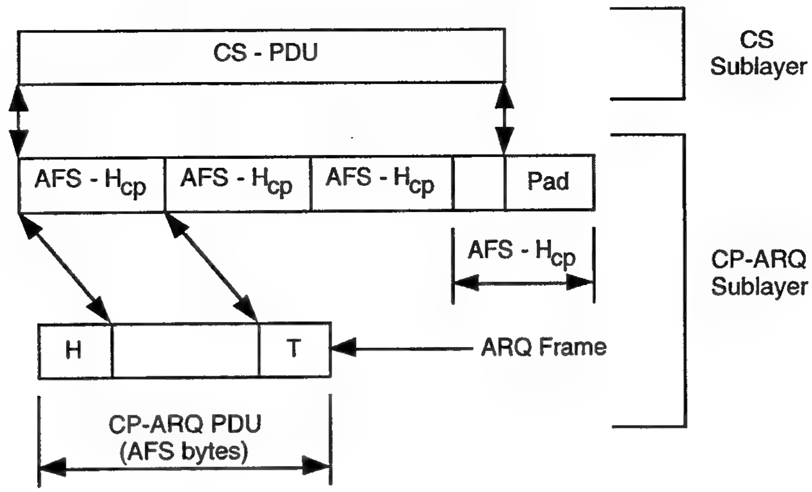
Table 3: AFS Factor Lookup Table for Host C

In a network where Switched Virtual Circuits (SVCs) are in use, the established link's AFS Factor can not be pre-determined by the system administrator, because SVCs are established dynamically by the network upon request of a user. In this case, a software routine should be developed that calculates this value and places it into the tables.

Additionally, if the host was going to be using static point-to-point links, and performance was an issue, the AFS Factor could be coded directly into the operating system kernel driver, thus requiring no lookup overhead from a table. This would be a flexibility vs. performance decision.

Since the CP-ARQ SDU will not necessarily be a multiple of 48 bytes, padding will be necessary. *Figure 6: AALx Padding*, shows where the pad is placed on the CP-ARQ SDU and how the CP-ARQ sublayer will calculate the amount of padding to add.

Determination of the **PAD** Field Size



where,

$$\text{Pad} = (\text{AFS} - H_{cp}) - (\text{CS-PDU} \ \% \ (\text{AFS} - H_{cp})) \text{ bytes}$$

and

$$H_{cp} = H + T \text{ bytes}$$

Figure 6: AALx Padding

The figure shows a single ARQ Frame at the bottom. The size of the ARQ Frame is ARQ Frame Size bytes (AFS bytes). Since there are H_{cp} bytes of header and trailer information, then the CP-ARQ PDU payload is reduced to AFS - H_{cp} bytes. To determine the size of the pad field, the size of the non-shaded portion of the CP-ARQ SDU next to the shaded box labeled *Pad* must be determined. The part of the pad equation that reads

$$(\text{CS-PDU} \ \% \ (\text{AFS} - H_{cp})),$$

determines that amount. The percent sign (%) represents the modulo function. The modulo function returns the remainder of the division between two values. That result is then subtracted from the number of bytes that represents the CP-ARQ PDU payload.

A note on efficiency. Even though padding only exists on the last ARQ frame, it still impacts the efficiency. To minimize the impact of padding on the efficiency, a lower AFS should be chosen. But, a larger AFS increases efficiency. The impact of the AFS is larger than the impact of the padding. Here lies a design constraint. Experiments similar to those conducted by Harris Corporation should be conducted to determine the maximal AFS.

We have, thus far, covered the choice of placement of the CP-ARQ sublayer, the choice of the AALx-SAR and AALx-CS sublayers, the method by which the size of the ARQ frame will be determined, and the padding algorithm. We have yet to discuss

which ARQ mechanism we have chosen, the size of the sequence number space, or which error detection method has been chosen.

Next the ARQ mechanism. Since AALx is being designed for use in the tactical environment, where there are a wide variety of tactical links available, the logical conclusion is to choose the Selective Repeat ARQ. The quantitative results prove that the Selective Repeat ARQ is able to perform better over a wider range of link characteristics than is the go-back-n ARQ.

In order to design the sequence number space, is was necessary to find the practical worst case tactical link bandwidth-delay products (remember the delay is worst case round-trip delay). For line-of-sight (LOS) radios (TSSR - Troposcatter Satellite Support Radio) the pair of 45 Mbps and 0.25 seconds resulted. Calculation of the bandwidth-delay product yields 11.25 Mbits worth of information. In order to achieve maximum efficiency, the bit-pipe must be filled at all times. A sequence number field of 24 bits is able to cover 16 Mbits. Therefore both the sequence number (SN) and request number (RN) fields will be 3 bytes, for a total of 6 bytes. Now, we know that each frame is not a single bit, so do we actually need all of these bits for the sequence numbers? The answer is no, we don't, but since the ARQ frame size is adaptive we can not calculate the sequence number space exactly.

As mentioned earlier, the window size, n, is the result of the bandwidth-delay product divided by the ARQ frame size. Since the AFS is variable, so must n be. Because of the window size's dependency on the AFS, a logical method to assign *n* would be to take the lookup tables, Tables 2 & 3, and expand them to include *n*, as in *Table 4: CP-ARQ Lookup Table w/ Window Size.* As for the AFS, software routines should be made to calculate the window size as well.

Host	n	AFS Factor
A	1024	256
B	512	512
D	512	32

Table 4: CP-ARQ Lookup Table w/ Window Size

On to error detection. A parity check, or a cyclic redundancy check? Well, for reasons discussed briefly above, a cyclic redundancy check provides much better coverage than a parity check for the same field size. Additionally, CRCs have been

used widely and successfully in protocols for a long time. Therefore, a 16-bit CRC has been chosen to perform the error detection.

Before we conclude, an important issue must be covered: multiplexing. The AAL3/4 SAR sublayer has one additional 10-bit header field called the multiplexing identifier (MID) field. With this field, up to 2^{10} user to user CS connections can be multiplexed across a single ATM user to user connection. Multiplexing is not always supported in AAL3/4 implementations, but AALx should be able to support multiplexing. The current design of AALx will only allow for a single user to user CS connection across a single user to user ATM connection. Because of this, a second AALx design was made that took multiplexing into account.

Recalling that the AALx SAR sublayer is identical to the AAL5 SAR sublayer. By using the AAL5-SAR sublayer, the functionality of multiplexing is lost. In order to support multiplexing, the AALx CP-ARQ PDU as well as the AALx SAR PDU structure had to change. These changes are shown in *Figure 7: AALx CP-ARQ Frame Structure w/ MID Support* & *Figure 8: AALx SAR Frame Structure w/ MID Support.*

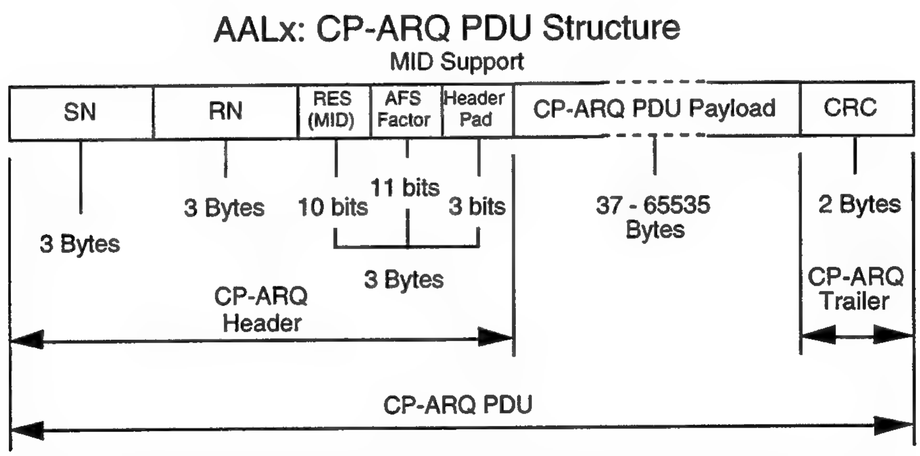


Figure 7: AALx CP-ARQ Frame Structure w/ MID Support

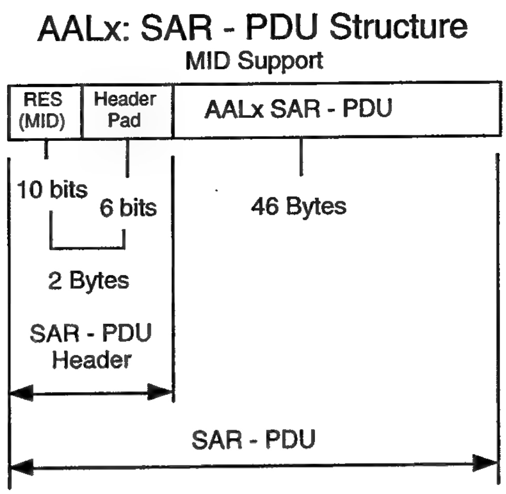


Figure 8: AALx SAR Frame Structure w/ MID Support

The PDU structure changes are summarized as follows: the CP-ARQ header requires an additional byte of information, the SAR requires 2-bytes of header information. As for the other aspects of the design, nothing has changed. When determining the pad for the CP-ARQ SDU, use a value of 11 for the H_{CP}. When determining AFS $Factor_{max}$, the result was 1425, which is still covered by 11 bits. To calculate the AFS, use

$$(\text{AFS Factor}) * (46).$$

<u>Conclusions</u>

The AAL design presented in this paper is a proposed solution to the problem of ATM in the tactical environment. This design will provide TCP with the segment stream it needs to provide quality-critical services to applications. There are still many issues that still need to be addressed. How will software routines determine the link characteristics and build the lookup table? Should multiplexing be used? Simulations need to be run for this protocol stack to determine the protocol parameters (AFS and n) that yield the best performance. To further tailor the protocol, exact values for the tactical communication links should be used in the equations presented here to allow the best results. Overall, this design is hope for TCP. It may be able to be used successfully in the tactical environment. Only simulations can tell.

References

Alles, Anthony [1992], *Tutorial: ATM in Private Networking.* Huges LAN Systems. Mountain View, California.

Apple Computer, Bellcore, Sun Microsystems, and Xerox [October 1992], *Network Compatible ATM for Local Network Applications: Phase I.*

Cassidy, P. [May 31, 1995], *ATM Information Transfer Over Degraded Comm Links*, Joint Advanced Demonstration Environment (JADE) Review, Technology Coordination Meeting of the JDL Communications Networks Subpanel.

Comer, Douglas E. [1995], *Internetworking with TCP/IP: Volume I: Principles, Protocols, and Architecture. Third Edition.* Prentice-Hall, Englewood Cliffs, New Jersey.

De Prycker, Martin [1995], *Asynchronous Transfer Mode: Solution for Broadband ISDN.* Prentice-Hall, Englewood Cliffs, New Jersey.

Gallager, Robert, and Dimitri Bertsekas [1987], *Data Networks.* Prentice-Hall, Englewood Cliffs, New Jersey.

Harris Government Aerospace Systems Division [May 1995], *Performance Analysis of ATM Networks with Wireless Links*, Melbourne, Florida.

International Telegraph and Telephone Consultative Committee (CCITT), Study Group XVIII, Geneva. [June 9-19,1992], *I.363 - B–ISDN Adaptation Layer (AAL) Specification.*

Le Boudee, Jean-Yves [1992], The Asynchronous Transfer Mode: a tutorial. *Computer Networks and ISDN Systems*, 24, 279-309. North-Holland.

McKeown, Nick [April 1994], *ATM: From A to B.* Lecture Notes.

Merriam-Webster, A [1973], *WEBSTER's New Collegiate Dictionary.* G. & C. Merriam Company, Springfield, Massachusetts, U.S.A.

Onvural, Raif O. [1995], *Asynchronous Transfer Mode Networks: Performance Issues (2nd edition).* Artech House, Norwood, Massachusetts.

Scholz, J.B. [1995], *Retransmission Operation of TCP/IP Over Radio Links.*

Scholz, J.B., and P. Cassidy [January 1995], *The Operation of TCP and UDP Protocols Over ATM Radio Links.* DSTO, Australia and Rome Laboratory, New York.

Scholz, J.B., P. Cassidy, M. Draper, and S. Lanigan [1995], *Experiments to Examine the Performance of ATM over Tactical Communications Links.* DSTO, Australia, Rome Laborotory, New York, and Rome Research, New York.

Scholz, J.B., P. Cassidy, M. Draper, S. Lanigan, and B. Millar [1995], *ATM Information Transport Over Degraded Communications Links.* DSTO, Australia, Rome Laborotory, New York, and Rome Research, New York.

Scholz, J.B., P. Cassidy, S. Lanigan, and M. Draper [1995], *Experimental Investigation of Multimedia Communications Over Degraded ATM Links.* DSTO, Australia, Rome Laborotory, New York, and Rome Research, New York.

Stevens, W. Richard [1994], *TCP/IP Illustrated, Volume 1: The Protocols.* Addison-Wesly, Reading, Massachusetts.

AIRBORNE BISTATIC CLUTTER MEASUREMENTS: SYSTEMS ISSUES

Elizabeth M. Twarog
Graduate Research Assistant
Department of Electrical and Computer Engineering

Northeastern University
309 Dana Research
Boston MA 02115

Final Report for:
Graduate Student Research Program
Rome Laboratory Hanscom

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory Hanscom

September 1996

AIRBORNE BISTATIC CLUTTER MEASUREMENTS: SYSTEMS ISSUES

Elizabeth M. Twarog
Graduate Research Assistant
Department of Electrical and Computer Engineering
Northeastern University

Abstract

This paper reviews airborne bistatic radar scattering programs that have been reported in the open literature. Brief descriptions of the systems considerations of these programs, with respect to hardware, synchronization and calibration issues, and experiment geometry are given. A preliminary design of an airborne bistatic measurement program is given, based on a potential transmitter and receiver. This report summarizes the author's research performed over a 30 day period during the summer of 1996 as a participant in the AFOSR Graduate Student Research Program at Rome Laboratory, Hanscom AFB, MA.

AIRBORNE BISTATIC CLUTTER MEASUREMENTS: SYSTEMS ISSUES

Elizabeth M. Twarog

I. Introduction and Motivation for Airborne Bistatic Clutter Measurements

There is a lack of bistatic clutter measurements at high grazing angles. Obtaining clutter measurements using a ground-based radar system at grazing angles above a few degrees with baselines longer than tens of meters would require prohibitively high towers or such platforms. This introduces the added complexity of an airborne platform into an already complicated bistatic scenario.

Some technical difficulties associated with an airborne bistatic collection program are the problems of synchronization of the transmitter and receiver, the complex clutter cell geometries associated with transmitter and receiver beam overlap, and the concern with overall system calibration to produce valid normalized radar cross section (NRCS) values. Introducing an airborne transmitter and/or receiver adds motion and position uncertainties along with a drastic increase in expense.

This level of complexity of airborne bistatic measurements is reflected in the small number of available published studies in the open literature. Five such studies were found and summarized here with respect to their hardware and experiment geometry and other considerations including calibration, synchronization, and transmitter/receiver beam overlap. These five studies were the SRC/AMBIS program, the ERIM HBR program, the ERIM/DARPA bistatic measurement program, JHU/APL bistatic measurements of the sea, and the Bistatic-MCARM program. Two of these five programs presented no clutter measurements.

Using the lessons learned from these studies, the final section of this report presents a preliminary design of an airborne bistatic clutter measurement system. With the idea of utilizing existing hardware in order to keep costs low, a possible transmitter, the NRL X-band RAR, and an available NU/RL S-band receiver are examined for the

feasibility of such a program. Issues such as hardware modifications, SNR, possible geometries or flight plans, system calibration, and synchronization are discussed.

II. Review of Airborne Measurement Programs Planned and Conducted to Date

A preliminary step in the design of any experiment is to examine any existing reports dealing with similar types of experimental programs that have already been conducted. A search of the open unclassified literature produces very little information on bistatic scattering; even less on air-to-air or air-to-ground experimental programs. Of the five program descriptions that were possible to locate, only three actually produced calibrated cross sectional values. In each of the three programs with data, only a single linear polarization was transmitted. One of the two reports without experimental data was a test plan report from 1995 with plans to conduct the described experiment in the months following the report publication. The other report without data was canceled due to funding and indicates no plan to follow through with their experimental design. The following section summarizes the five experimental plans. Operating parameters for all experiments are given at the end of the section in Table 1.

A. AMBIS (Adaptive Multimode Bistatics):

The Cudjoe Key Bistatic Clutter Testing Program [1] was performed during 10-13 November 1994 and was composed of a tethered aerostat-borne S-band transmitter stationed at Cudjoe Key, Florida, and an airborne receiver carried in a Cessna. The aircraft flew north-south and northwest-southeast racetrack patterns over the Gulf of Mexico and Florida Straits. Collected terrain data consisted of gulf water (wave heights 0'-4'), ocean (wave heights 4'-8'), everglades, and keys. Transmit and receive grazing angles varied from 1 to 17 degrees with out-of plane angles from 20 to 135 degrees.

The transmitter SEEK SKYHOOK was a DPS-5 radar, tethered to a tower at Cudjoe Key, Florida. The nominal operating height was 10,000 ft. The mobile receiver was installed in a Cessna 402B aircraft. Two L/S-band horn antennas were mounted in the aircraft nose, and a third was located inside the aircraft, pointing out the starboard

window. During this experiment, the AMBIS system used wide receive beams to cover the region where the transmitter would be flying, instead of using pulse-chasing techniques for larger area surveillance. An on-board GPS receiver was used to time-mark the data.

This system utilizes a non-cooperative transmitter; i.e. the receiver exerts no direct control over the transmitter. With no dedicated line-of-sight link from the transmitter, the receiver must derive the necessary synchronization and coherent signal information by intercepting a line-of-sight signal from the transmitter. No atomic clock synchronization was used. The intercepted transmitted waveform is used as a coherent processing reference which is cross-correlated with the received scattered signal to remove phase uncertainties and provide a zero-time-delay reference for the data.

Typically, a separate reference channel in the receiver records direct path and short time delay multipath signals. However, for this experiment, to avoid loss of continuous range gate sampling after the direct path signal during the interval between switching from data collection in reference and data channels, a single channel of data was collected and a "synthetic" reference pulse was derived from an unsaturated sample of the direct path signal.

Noise data was collected at the beginning and end of each flight. After data collection, 16 pulses of the data are non-coherently integrated to get an estimate of the clutter-to-noise ratio for each sample. Direct path calibration was performed by maneuvering the plane so that each of the three receive horns collects the direct path signal in turn. Theoretical and measured direct path signals were then compared. The GPS time-stamp is used to derive the bistatic geometry of the clutter cells needed for calculation of σ_0 values. By knowing the transmitter and receiver positions, ranges and angles to the clutter cell can be determined, and from that, the clutter cell area and antenna pattern attenuation.

Measured cross sections with clutter-to-noise ratio values greater than 10 dB are reported for the four terrain groups of gulf, ocean, everglades, and keys. Values of σ_0 are presented as functions of out-of-plane angle, transmitter grazing angle, and receiver

grazing angle. The authors estimate an accuracy of 3.7 dB in σ_0 values based on the estimation accuracy of the clutter-to-noise ratio as well as system and clutter cell geometry parameter estimation. No analysis of data trends was presented in this report.

B. ERIM HBR (Hybrid Bistatic Radar) Program:

The Hybrid Bistatic Radar program [2] was aimed at determining the viability of utilizing the bistatic enhancement of low RCS airborne targets for target detection. Their published final report is a system study and experiment design that never progressed past the design phase, so no measurements were performed. The primary contribution of this report was the discussion of the calibration concepts, along with details and test results of the active radar calibrator (ARC) that was built to resolve those problems. While this report provided no measurement data, it did detail many important considerations necessary when designing an airborne bistatic clutter measurement program. It should be a useful guide to other interested researchers.

Clutter measurements were to be made from two airborne platforms flying on parallel courses. Such a geometry results in ground clutter measurements at large bistatic angles and shallow grazing angles. Interleaving monostatic and bistatic scattering data on one receiver would allow for essentially simultaneous cross section results.

The final collection system design uses ERIM's two L-band transmitters installed on two Convair CV-580 aircraft. Separate monostatic and bistatic transmitters and a single common receiver would be used, as two transmitters would be cheaper and less complex than two receivers. A real-time data link would be used to maintain correct aircraft position, with atomic clocks for system coherence and PRF timing.

An active radar calibrator is needed to calibrate a bistatic system in order to provide a large cross section in the presence of high background clutter cross sections. Passive calibration targets with large enough cross sectional values would have impractically large physical dimensions. One difficulty with this type of active bistatic calibrator is that multipath effects will cause a decrease in accuracy at low grazing angles. To combat this problem, ERIM proposed to calibrate the system at grazing angles above

20 degrees and then transfer the calibration down to the lower grazing angles that would be used in the experiment. The ERIM ARC consists of separate receiving and transmitting antennas and includes both a time delay (400 ns) and a Doppler offset (0-250 Hz) in order to place the calibration signal in a time-Doppler cell without predominant background clutter levels. Field tests of the ARC indicated that it would perform as expected.

C. ERIM Measurements of Bistatic Cross Section:

A DARPA-sponsored bistatic air-to-ground program was conducted by ERIM in two published reports [3,4]. Both phases of the program utilized a dual L/X-band airborne transmitter flown in a C-46 aircraft and a tower-mounted receiver. Each was scanned to spotlight the terrain. The first published paper, covering the period 1 October 1975 to 30 September 1976, presented returns from two terrain types: smooth concrete with grass, and dense, tall, dry weeds. Incidence and scattering angles ranged from 50 to 80 degrees and out-of-plane angles from 0 to 180 degrees. The second published report, covering the period 1 September 1977 to 30 June 1978, presented three terrain types: rough ground with dry weeds with a 6 inch snow cover, rough ground with dry weeds with a 12 - 20 inch snow cover, and a tree-covered site with a 6 inch snow cover. Incidence and scattering angles ranged from 60 to 84 degrees, with out-of-plane angles again between 0 and 180 degrees.

The two sets of measurements used the same air-to-ground system. A dual L/X-band transmitter was flown in ERIM's C-46 aircraft. Both wavelengths were transmitted simultaneously, using horizontal polarization only. The receiver was mounted on a ground tower and collected both co- and cross-polarized returns. As the aircraft flew along the flight line, the transmitting antennas were scanned to spotlight the scattering area. The receiving antenna is also scanned, at a rate of one scan as the transmitting antenna illumination angle changes by 5 degrees. The aircraft is tracked along each line and the bistatic angle of the scattering patch is determined by the difference between the recorded transmitted and received nominal illumination angles.

The receiver is synchronized to the transmitter upon reception of the direct path signal from the transmitter. A variable delay in the receiver then turns on the receiver, so in-plane forward scattered signals are not received. Coherent data is sampled on each of the four channels on a pulse-by-pulse basis. Received power is averaged over 5 degree intervals of azimuth angle.

Both the paper and the report present cross section values versus out-of-plane angle for the various incidence and scattering angles. For the data presented in the paper, X-band cross sections as low as -20 dB were measured and were always higher than the L-band cross sections, which were as low as -40 dB. A more detailed analysis of the second experiment was presented in the report. The reported error estimate was 4 or 6 dB, depending upon the month in which the data was collected. Comparison of the results from the 1976 and 1978 experiments at the same site showed little variation in the cross-sectional values at L-band, since the snow cover is much less than the penetration depth at L-band. Uncertainties in the X-band data were too large for comparison between the two sets.

D. Bistatic Cross Section of the Sea - JHU/APL:

Bistatic measurements of the sea at three different sea states were made by the Applied Physics Lab at Johns Hopkins Lab in the early 1960's [5]. A land based C-band CW transmitter was used, along with a receiver flown in an aircraft. For each flight, the receiver aircraft flew directly over the transmitter and out to sea. Data was collected at transmitter grazing angles of 0.2 to 3 degrees with receiver depression angles from 10 to 90 degrees. All scattering was in-plane.

Limited details about the radar hardware and logistics of the experiment were published in the communication. The transmitter was a CW C-band system located on the shore which transmitted vertical polarization only. The elevation angle was fixed, but the antenna could be manually positioned in azimuth to illuminate the sea below the flight path. A second radar was used to track the receiver aircraft during the experiment. The airborne receiver collected both horizontally and vertically polarized returns. The

receiver portion of the air-to-ground system was calibrated by flying the receiver towards the transmitter and using a narrow transmitter beam with a fixed flight altitude. With all geometry parameters known, absolute power received can be calculated.

Co- and cross-polarized σ_0 values are presented for sea states 1, 2, and 3, corresponding to wind velocities of 5, 20-30, and 10 knots, respectively. The authors note that the wind velocities do not correspond to the expected winds for a given sea state and a fully arisen sea, and remark that the wave heights were measured by a spar buoy anchored a mile off shore near the illuminated area. The data shows that cross sectional values increase with transmitter depression angle. Cross-polarized values were lower than co-polarized values for all values of transmitter depression angles but increased more rapidly with depression angle than co-polarized values. The σ_0 values are relatively independent of receiver depression angles between 10 and 90 degrees. NRCS does show a dependence on sea state, implying a dependence on the wind over an extended area.

E. Bistatic Multi-Channel Airborne Radar Measurement (Bistatic MCARM) Data Collection Program:

The Bistatic MCARM program [6] was interested in high PRF data at small bistatic or pseudo-monostatic geometries. The objectives of this program was to develop a bistatic clutter base for use in space-time processing algorithms and in bistatic clutter models to estimate the impact of hot clutter. The test program uses a Tethered Aerostat Radar System (TARS) at Horseshoe Beach, Florida, as a transmitter and the MCARM system in a receive-only mode as the receiver. As of the publication date of this report, 21 April, 1995, the flights described had not taken place. This test plan, therefore, should be used as a guide for an experimental set-up in the same way as was indicated in the ERIM HBR program.

The TARS transmitter will be tethered at a height of 10 to 15 kft, providing line-of-sight range out to a horizon beyond 75 nmi. It is an L-band surveillance radar which transmits vertical polarization using an elliptical (25' by 13') rotating reflector. The

bistatic receiver will be flown on a BAC 1-11 aircraft at an altitude of 10 kft. The receiver STALO/waveform generator was replicated to be the TARS signal source, eliminates the problem of locking onto a non-cooperative transmitter. Aircraft GPS and INS data will be recorded at the receiver.

Also included in this test plan is a multi-target simulator (MTS), which will serve as a field located calibrated target and can emulate targets of varying Doppler and amplitude characteristics. The terrain near the Horseshoe Beach site includes the water of the Gulf of Mexico, marshes, inland pine trees, and some urban clutter. The proposed flight path consists of 5 straight legs with bistatic baselines varying from 25 to 86 nmi.

III. Preliminary Design of an Airborne Bistatic Measurement System

The previous section summarized the methods that other researchers have used to solve the problems of calibration, synchronization, and calculation of clutter cell geometry. This section outlines a preliminary design of an airborne bistatic clutter measurement system, presenting a possible airborne transmitter or opportunity, and an existing NU/RL ground-based receiver. An air-to-air collection scheme is not considered because of the complexity of obtaining and scheduling two airborne platforms. A first-order signal-to-noise analysis of the two systems in a bistatic configuration is presented. A discussion of geometry, synchronization, and calibration follows.

A possible airborne transmitter is the NRL X-band real-aperture radar (RAR). It operates at 9.375 GHz, using a 20 kW magnetron and a 3m dual-polarized slotted waveguide antenna. The antenna has a 3dB azimuth beamwidth of 0.9° and elevation beamwidth of 22°. System characteristics are summarized in Table 2. The RAR is installed looking out the port side of an NRL P-3 aircraft and can be switched between 45°, 60°, and 80° incidence angles.

Table 1. System Parameters for the Five Experiments

Experiment	Frequency	Platform	Polarization	Pulsewidth	Beamwidth	Data	Grazing Angle	OOP Angle
AMBIS	S-band 3.23 GHz	Tethered Tx, Airborne Rx	V Tx V Rx	1.25 ms	1.75° Tx 25° Rx	gulf, ocean, everglades, keys	1 - 17°	20 - 135°
ERIM HBR	L-band 1.6 GHz	Proposed airborne Tx, Rx. Canceled due to funding	4 linear pol	125 ns	6°az, 17°el Tx and Rx	none	-----	-----
ERIM/DARPA	X-band 10 GHz; L-band 1.3 GHz	Airborne Tx, Tower-based Rx	H Tx V, H Rx	(1)	(1)	concrete, dry weeds, ground + snow, trees + snow	50 - 84° (incidence)	0 - 180°
JHU/APL	C-band	Fixed Tx, Airborne Rx	(2)	(2)	(2)	Sea states 1-3	0.2 - 3° Tx 10 - 90° Rx	in-plane
MCARM	L-band 1.24-1.27 GHz	Tethered Tx, Airborne Rx	V Tx V Rx	1.6 - 100 us	2.2° Tx 11.1° az, 11.9° el Rx	none	-----	-----

(1) Unable to find instrument documentation published in a separate report

(2) Not reported in published communication

Table 2: X-band RAR System Characteristics

Frequency	9.375 GHz
Power	20 kW
Pulse Repetition Frequency	100 Hz
Elevation 3dB Beamwidth	22°
Azimuth 3dB Beamwidth	0.9°
Pulse Width	50 ns
Gain	33dB

An available receiver is the NU/RL S-band dual-polarized mobile receiver [7]. It operates at 2.71 GHz with an intermediate frequency (IF) of 60 MHz. The antenna would be a wideband standard gain horn with a 20 dB gain. This receiver could be installed in a van and deployed in the field. An obvious problem with using this receiver with the NRL RAR is the discrepancy in the operating frequencies. Some modifications would have to be done to the receiver. One potential solution would be to us an X-band horn antenna, mix the signal down to S-band, then use the rest of the S-band receiver without further modifications. A second solution would be to remove the entire S-band front end and replace everything with X-band components down to the local oscillator mixer. The practicality and cost of each modification warrants further examination. The use of this proposed system would allow for the collection of VV, HH, and cross-polarized NRCS values, something not produced in any of the published studies to date.

The SNR for the total system must be calculated in order to determine the feasibility of using the NU/RL receiver and NRL transmitter. The system parameters from Table 2 were used and the following assumptions were made.

As a first approximation, the monostatic signal-to-noise can be calculated and then degraded by the difference in gains of the transmitter and receiver. This is assuming that the ranges from the transmitter and receiver to the clutter patch is the same, along with such system parameters as the noise figure and internal losses. If a large floodbeam

receiver is used, the clutter cell geometry can be approximately determined by the azimuth angle and pulsewidth of the transmitter.

The monostatic cell area is given by $A \cong R\Theta(c\tau/2) \cong 7R$. The latest estimates of bistatic low grazing angle terrain clutter NRCS values range from -30 to -50 dB. Using an NRCS of -40 dB, the cell RCS is $\sigma = \sigma_0 A = 7*10^{-4}$.

The signal-to-noise ratio is calculated by:

$$SNR = P_t G_t G_r \lambda^2 L_s \sigma / (R^4 (4\pi)^3 K T_o B F_n) = 129 - 3R$$

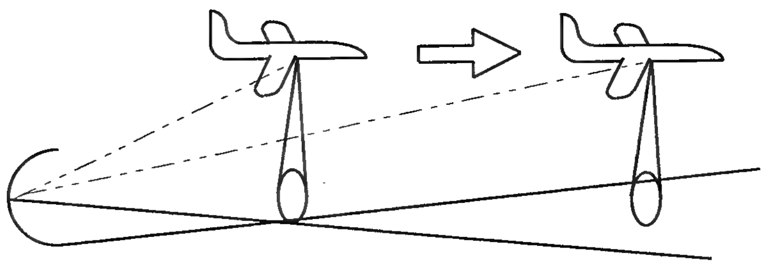
where the following assumptions were made:

$B = 2/\tau$ $\qquad$ $F_n = 5$ dB

$L_s = -5$ dB $\qquad$ $G_t = 4\pi/(\theta az)(\theta el)$

Using a range of 1000m, the SNR would be 39 dB. This value is adequate for clutter measurements. Even at an altitude of 3 km or with a lower clutter RCS, the SNR will be acceptable. This first order type of approximation indicates that a clutter measurement program using a transmitter with a low level of complexity such as the 20 kW magnetron would be of reasonable size, along with the antennas for both the transmitter and receiver. Once a particular bistatic collection scenario is designed, a more thorough analysis of the SNR must be completed.

In bistatic radar, the range cell areas are a more complicated function of geometry than in a monostatic configuration, forming the so-called ovals of Cassini. When one of the platforms, either the transmitter or receiver, is put in motion the geometry becomes further complicated, as the bistatic baseline distance and/or position changes. For example, in the simple scenario in the figure below, the bistatic baseline, as indicated by the dash-dot line, lengthens as the aircraft bearing one antenna flies away from the other antenna.



In this scenario, the range ovals would become increasingly elongated. For a more complicated flight path, such as when the aircraft crosses the beam perpendicularly, the skewness of the ovals would change with time. Such a flight plan is illustrated in the figures below. In these three figures, one ground-based antenna and one aircraft-borne antenna is assumed. The aircraft was chosen to be flying at 400m.

The proposed NU/NRL S-band transmitter does not have the capability to be quickly scanned in order to perform pulse chasing, so to maximize the area covered by the bistatic system, a floodbeam horn should be used. The transmitter can be flown in two bistatic scenarios: along the receiver beam axis, or into the receiver beam, crossing the beam axis at some angle, and then out of the beam. Flying along the beam would be easiest to coordinate, but would yield only in-plane scattering angles. Crossing through the beam could produce a variety of both scattering and bistatic angles.

One way the AMBIS program solved the problem of determining the clutter cell geometry was to time stamp the data with a GPS mark. If the time and position of the transmitter are known, the clutter area can be computed. Our stationary receiver will slightly simplify this problem by eliminating position and aspect uncertainties inherent in the airborne receiver. A possible method of synchronizing the transmitter and receiver, following the ERIM report [4], would be to synchronize the receiver upon reception of the direct path signal from the transmitter. Data collection would then begin at a specified time delay.

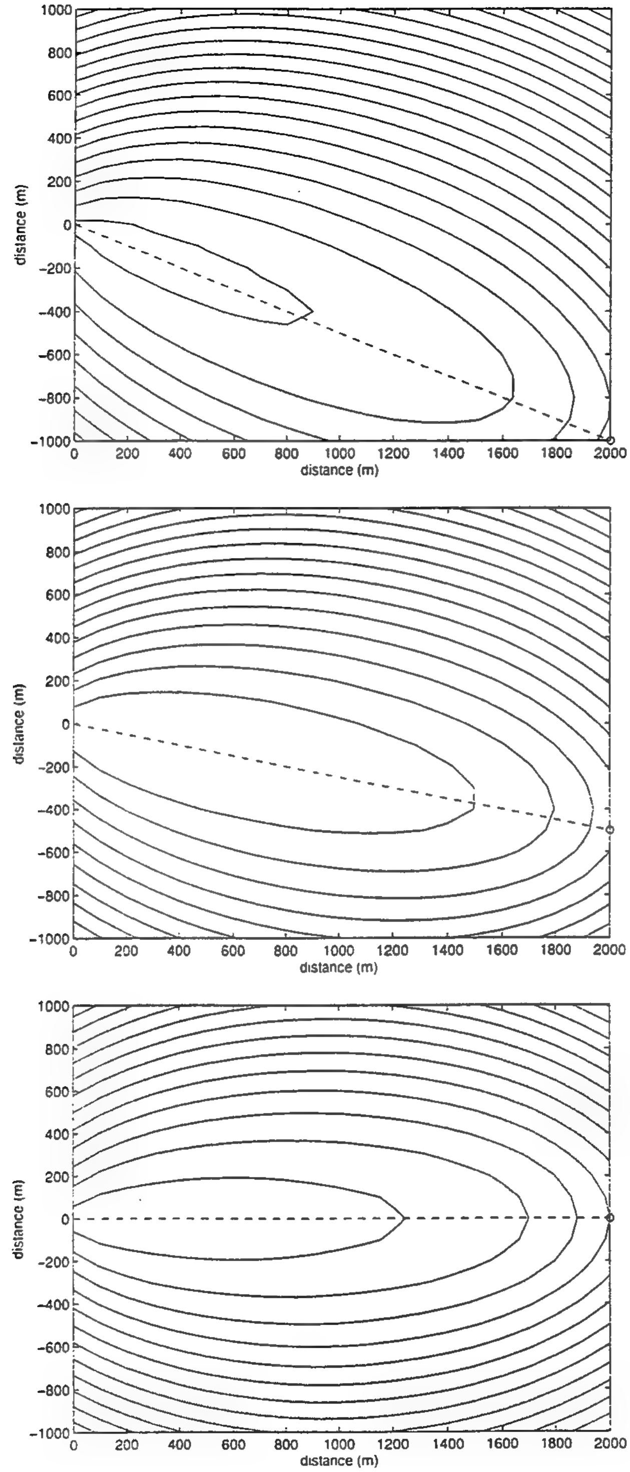


Figure __: Range contours as aircraft flies in +y direction. Aircraft height is 400m and aircraft position is marked by a circle on the far right side. Dotted line indicated bistatic baseline. For all three figures, the contour spacing is 100m. The first contour for the three figures is at 2300m (top), 2100m (middle), and 2000m (bottom).

The final major consideration is the calibration of the system. If the X-band RAR is used as the transmitter, the magnetron oscillator will result in incoherent measurements only. In this case, a direct path measurement link between the transmitter and receiver will suffice for calibration. This could be accomplished by flying a flight track which perpendicularly crosses the receiver beam axis. If fully polarimetric measurements are required, another transmitter must be used and the system could be calibrated using a polarization-agile beacon, as described in McLaughlin et al.

IV. Conclusions

This report summarized the study of all available reports in the open literature dealing with air-to-air or air-to-ground bistatic measurement program. Five such programs were found, and three of the five reported calibrated NRCS values. This study was to detail the considerations for such a potential study. The preliminary design for an airborne bistatic measurement study using a transmitter of opportunity and an existing NU/RL receiver was presented and the SNR for such a configuration was shown to be adequate. Calibration, synchronization, and clutter cell geometry were discussed.

V. References

[1] Syracuse Research Corporation, "Bistatic Clutter Measurement Program, Phase 3", Prime Contract Number: F30602-93-D-0081, May 16, 1995.

[2] S.J. Anthony, et al. ERIM, "Hybrid Bistatic Radar Clutter Measurements Program", Contract F30602-86-C-0055.

[3] R.W. Larson, A. Maffett, F. Smith, R.C. Heimiller, and A. Fromm, "Measurements of Bistatic Clutter Cross Sections", RADC-TR-79-15, May 1979.

[4] R.W. Larson, A.L. Maffett, R.C. Heimiller, A.F. Fromm, E.L. Johansen, R.F. Rawson, and F. Smith, "Bistatic Clutter Measurements", *IEEE Trans. Antennas Propag.*, vol. AP-26, pp. 801-804, 1978.

[5] V.W. Pidgeon, "Bistatic Cross Section of the Sea", *IEEE Trans. Antennas Propag.*, vol. AP-14, pp. 405-406, 1966.

[6] Westinghouse Electronics Systems Group, Baltimore, MD, "Bistatic Multi-Channel Airborne Radar Measurement (Bistatic-MCARM) Data Collection Program Test Plan, for Rome Laboratory, 21 April, 1995.

[7] D.J. McLaughlin, R.S. Raghavan, Y. Wu, and N. Pulsone, "A Study of Polarimetric Strategies for Suppressing Bistatic Clutter with Applications to Radar Detection", USAF Rome Laboratory Contract No. F19628-93-K-0005, Interim Technical Report, August 1994.

[8] D.J. McLaughlin, Z. Ren and Y. Wu, "A Bistatic Polarimeter Calibration Technique", *IEEE Transactions Geosci. Remote Sensing*, Vol. 33, No. 3, pp. 796-799, 1995.

INCORPORATING AN HPC PARALLEL TRACKING PROGRAM INTO A DISTRIBUTED, REAL-TIME, TRACKING APPLICATION.

Phillip W. Young
Master's Student
Department of Computer Science

University of Connecticut
U-155
Storrs, CT 06269

Final Report for:
Graduate Student Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

August 1996

INCORPORATING AN HPC PARALLEL TRACKING PROGRAM INTO A DISTRIBUTED, REAL-TIME, TRACKING APPLICATION.

Phillip W. Young
Master's Student
Department of Computer Science
University of Connecticut

Abstract

A parallel tracking program, running on an Intel Paragon high-performance computer (HPC) has been incorporated into a real-time, distributed application. The tracking program was merged into the "Multisource Integrated Software Tool" (MIST v2.0) developed by Rome Laboratories. MIST was created to provide an environment in which various target tracking methodologies could be developed and compared. To that end, TCP sockets and XDR streams were added to MIST 2.0 in order to provide robust network communications between differing types of machines in an (inter)networked environment. A group of test program tools was also created to aid in the development of this and future tracking programs to be run on the HPC. Finally, the initial tracking program and final working system were analyzed. This analysis shows that the tracking module efficiently utilized the computing resources available. It also demonstrated that with sufficient computing resources the throughput of the system was limited by network communication rather than in tracking computation. The analysis also showed that, due to the reduction in time spent in the most computationally intensive portion of the tracer, it is now worthwhile to optimize the data association portion of the program.

INCORPORATING AN HPC PARALLEL TRACKING PROGRAM INTO A DISTRIBUTED, REAL-TIME, TRACKING APPLICATION

Phillip W. Young

Introduction

A multi-sensor multi-target tracking system should ideally be able to track a set of targets about which little information may be known. The tracker should be able to assimilate detections from targets of unknown number, and in the presence of noise, clutter, and missed detections.

In fact, much work has been done in the area of Multi-Sensor Data Fusion since the 1970's, focusing on the combination of data at the decision and track levels. Some of the better-known methods of performing this job are Joint Probability Data Association (JPDA), Multiple Hypothesis Tracking (MHT), and the pattern-recognition technologies of Track-Before-Detect, and Artificial Neural Networks. However, real-world constraints can have great impact on the performance of tracking approaches, and their suitability to a given 'environment'. In order to address this concern, MIST 2.0 was created to aid in developing and comparing various tracking methodologies using real and simulated data.

The target tracker which has been incorporated into the MIST 2.0 system employs an Interacting Multiple Model (IMM) *estimator* to estimate the positions of targets. By using multiple models to describe the behavior (such as landing or turning) of the targets during different time intervals, the IMM estimator tracks more accurately than a well-tuned Kalman filter [1]. The estimator provides some information about how likely it is that a particular detection originated from a particular track; this is the related problem of *data association*. The data association portion of the tracker is 'scan oriented'. The tracker assigns the latest scan (set of detections) to the track list (set of currently active tracks, as well as the possibility that a detection is from a new track), mapping this into a 2-Dimensional assignment problem. The 2-Dimensional assignment involves minimizing a global cost function of possible assignments

between the scan and track list. Of course, the set of detection-to-track associations that must be considered can be reduced, since some of these may be unrealistic. For example, one type of *gating* employed in an aircraft tracking scenario, known as *velocity gating*, allows the tracker to not consider associating a track with detections that are so far away from the last updated position of the track that the target would have had to move at an unreasonable velocity. Another method of reducing the amount of computation necessary to effectively track the targets under surveillance is *pruning*, i.e. removing tracks from the set of active tracks. Some tracks might terminate naturally or be initiated due to false detections. For example, in the case of the aircraft scenario, if a plane lands then its track need not be updated in future; or, if a cloud returns a spurious detection, initiating a new track, it would be inefficient to compute costs between this track and all subsequent detections. In both examples, the track will not likely be associated with any future detections, and so it may be dropped after some number of scans in which the track is not updated.

Although gating and pruning can reduce the number of association cost computations necessary, this task remains the computational bottleneck. Fortunately, these calculations can be carried out simultaneously [1]. This allows parallel computing to be applied, greatly reducing the amount of time spent in this portion of the tracker. In fact this is exactly what was done by Robert Popp [2] and his success in this area was the prime motivation in incorporating that parallel tracking program into the MIST 2.0 application.

The tracking program employs a supervisor-worker strategy in dividing the workload across the nodes available. Specifically, the set of tracks is distributed among the worker nodes such that no two nodes are responsible for the same track; and, the union of all worker node track sets is the set of all active tracks. Each node maintains the relatively large amount of data associated with the tracks for which it is responsible. For each scan of data, the node estimates the next position of each of its tracks and computes the costs of assigning each of its tracks to each of the detections. These costs are then sent by each of the worker nodes to the supervisor which assimilates them and determines the optimal assignment between the track list and the detections. The newly updated tracklist is then broadcast to all the worker nodes and the cycle repeats. The reader is referred to [1,3] for more information on the IMM tracker. figures 1 and 2

give a block diagram description of the general MIST 2.0 application. For more information on the Intel Paragon HPC the reader is referred to [2,4,5].

Discussion of Problem

A sequential version of the previously developed IMM tracking program was designed into MIST 2.0 which worked in a SUN UNIX environment. The parallel version of the IMM tracker was developed on the Paragon high performance computer. This version did not interface with MIST 2.0 and as such, the object of this work was to replace this sequential version with the parallel version. MIST 2.0 is structured in a client server architecture which is broken down into a data input process, a diagnostic process, a control & display process and the tracker process. The first three processes need to be run on a Sparc SUN workstation while the tracker process is replaced with the parallel version on the Paragon high performance computer. This resulted in a common MIST 2.0 architecture that can process real time data and a parallized tracker that runs on a high performance computer. Performing this effort entailed providing many solutions to various problems.

One problem was to integrate the IMM tracker's supervisor node into handling the data input and data output in an efficient manner. Since the supervisor processing node must wait for the worker nodes to compute their association costs, the tracking program used the supervisor node as a worker node. While this was an effective use of the available nodes in the stand-alone tracking program, which read all scans in before tracking, it required additional complexity and some redundancy in the tracking program. Most important however, it was necessary to provide a mechanism for the tracking program to communicate detections, tracks, and other information, to and from other networked machines working in the application, without destroying the efficiency of the tracking program. The tracking program was written as one process that was copied (forked) onto all the nodes in the application. These processes are synchronized using global synchronizing functions which required all nodes of the tracking application within the executing partition to participate. If communication (or other) processes were to be added to

other nodes within the *compute partition* of the HPC (where the tracking program runs) then they would be required to call the global synchronization functions as well. Because these functions are called in a sequence dependent on the data within the tracking program, these processes would have to process that data in a similar manner or use some inter-process communication mechanism in order to call these functions at the appropriate times. Both of these solutions would add a computational delay, and would create and unacceptable dependencies between these processes, making future development unnecessarily complicated.

Another problem that was solved was that simple data types (such as floating point numbers) are represented within the HPC and the Sparc machines in a different manner. While the problem needed to be overcome for these machines, a general solution to the (big Endian/little Endian) problem was required so that MIST 2.0 and its components could operate on (UNIX) machines of any type.

Also, at the time of the integration, MIST 2.0 itself was undergoing some development. The program employed UDP sockets and used signals frequently in its operation. The UDP sockets are connectionless - meaning that there is no guarantee that information will be receive in the sequence in which it was sent, or that it would be received at all. This seemed to pose no problem in machines on the local network in which MIST 2.0 was developed, but this was no guarantee in a larger (inter)networked environment. Similarly, signals, which allow a process to communicate the occurrence of event to another processes running on the same machine, are not always implemented with the same degree of reliability; so dependence on them was something to be reduced as much as possible. All efforts were made while maintaining the goal that the final system be robust in an (inter)networked environment, communicating amongst machines of various types, and that future work be as simple as possible to incorporate.

Methodology

Merging the tracking program hosted on the HPC into the MIST 2.0 hosted on the Sparc workstation required familiarity with both programs and both computer architectures. In performing this task, I consulted with Garry Fountain, Rome Reseach Corporation, who programmed the MIST 2.0, and Robert Popp, UCONN, who developed the parallel tracking

program. It was necessary to study both programs and to gain some level of proficiency with the HPC. In order to learn the programs, pseudo-code flow diagrams were created for the critical sections of each. These provided further insight into the operations of both programs and were useful in making design decisions later. In fact, it became apparent that some of the problems could share a common solution.

In order to provide reliable communication amongst the networked components of the program, TCP sockets were used to replace the UDP sockets. The TCP sockets provide a 'virtual-circuit', like a telephone connection, in which information is guaranteed to be received in the sequence in which it was sent, whereas the UDP sockets are 'connectionless'; that is, they don't even ensure that the information (packets) will all be received at all. The increased speed at the risk of information loss (using UDP sockets) might be acceptable in transmitting real-time audio or video data, but it wasn't considered wise in this application. Therefore, the UDP socket functions were replaced with a set of TCP functions. In testing the communications between the Sparc stations and the HPC, some small modifications needed to be made. A set of test programs was written so that the problem of integrating the two programs at a logical level and the problem of getting the HPC to operate as desired wouldn't overlap. These programs can be used in the future as starting points for integrating other programs on a hpc.

One set of four test programs (hpc_send, hpc_receive, sun_send, sun_receive) tested reading and writing data between processes that would be run on the Sparcs or the HPC in any combination. These programs were used to develop the XDR stream functions, used with the TCP streams, in order to provide a common representation of data types. The programs contain simple functions to test and demonstrate the use of XDR streams to and from files, between machines using memory buffering, and between machines by attaching to the TCP streams directly. It was decided that attaching the XDR streams directly to the TCP socket streams would be the most efficient way to proceed, since it obviated the need to declare temporary message buffer space or to require that the receiving program hard code the length of incoming messages; and because XDR streams can use the same function to encode and send a message as is used to receive and decode the message. Because communications between the *HPC-tracker* (the client finally incorporated into MIST 2.0) and the other components required that all communications

use the XDR streams, it was necessary to convert all network communication functions to use XDR streams. This was first done on MIST 2.0 using the existing tracking program running on Sparc-20 workstations. Once XDR streams were fully employed in MIST 2.0, the same communication functions could then be easily adapted into the HPC-tracker program. The next step was to modify the tracking program so that it would be compatible with the MIST 2.0 application.

The general structure of MIST 2.0 is that clients running on separate machines communicate through the network while subclients of each client handle incoming messages. The receiving client accepts a message, places it in a shared memory queue, and then signals the parent process that a new message has been received (a block diagram is shown in figures 1 and 2). In the case of the *Sparc-tracker* (the tracker originally running in MIST 2.0 on a Sparc workstation), a *detection client* receives incoming detections, and creates a *detection report* object, composed of the detections and some additional information. The detection client then stores the detection reports in a queue for later retrieval by the tracking client. As long as the queue holds at least one scan's worth of data by the time the tracker finishes processing the previous scan, the tracker process can work continuously. This technique allows the (tracking) process to work continuously, even if the data is received in bursts. The technique was necessary in the original implementation of MIST 2.0, which sent scans of data at fixed time intervals, regardless of the rest of the system. Since the tracker only requires one scan of data at a time, MIST 2.0 was modified so that the tracker would request detections whenever there were fewer than two detection reports in its queue.

Although most of the nodes on the HPC were running Paragon OSF/1, a subset of standard OSF/1 UNIX, and could employ shared memory, the HPC-tracker program does not use shared memory, since the HPC is a message passing oriented system. The detection client written for the HPC-tracker receives detections and broadcasts them to the supervisor and worker nodes upon request by the supervisor process. As mentioned before, the supervisor process was not as active as the worker processes, in fact 95% of the tracking period was spent waiting for the worker nodes to finish computing association costs [2]. It was for this reason that the supervisor was made into a supervisor/worker node in the previous tracking program; and this was also the

motivation behind the design of the HPC-tracking module incorporated into MIST 2.0. Since it would be unreasonable to add the communication delays to the tracking cycle, the supervisor was separated from the dual role of supervisor/worker so that the supervisor could transmit the previously calculated set of tracks to the display server while the worker nodes computed association costs. In order to avoid the problem of global synchronizations (mentioned earlier), the detection client process was placed on the same node as the supervisor process. Neither the Sparc-tracker nor the HPC-tracker actually used a tracking client to queue track lists for transmission to the display server, because the server queues the track lists as they are received. In order to be able to load different processes on different sets of nodes, it was necessary to create a *controlling process* on the HPC. Upon startup, the controlling process receives the system configuration over the network and loads various processes onto the nodes in the interactive compute partition of the HPC. The controlling process resides in the service partition, where it does not affect global synchronization function calls. Processes in the service partition, such as the controller, may receive little active time from the process scheduler. For this reason, the controlling process was only given the tasks performed by the Sparc-tracker's message client. Two test programs were created to create a 'generic, canned' controlling program and associated controlled programs. These were copied, modified, and developed into the HPC-tracking program. The same can easily be done for future HPC-tracking programs.

For performance comparisons, a function was written to quickly log the *actual* clock time at various points in the program. Data generated from these logs is given in the appendices and discussed below.

Results

The original MIST 2.0 hosted entirely on a Sparc machine is shown in Figure 1. Figure 2 shows the components used to interface to the hpc. Figure 3 shows how the tracker interface was modified to interface with the hpc. The HPC-tracking program structure and timing diagram are given in figure 4 and 6, respectively.

Figure 5 shows the performance of the original tracking program, which used a combined supervisor worker node, and compares it to the modified version, in which the supervisor process played a single role. It should be noted here that the times recorded in figure 5 are only those during which track formation related computations took place.

Figure 7 compares the performance of the HPC-tracking program employing varying numbers of nodes, against the original Sparc-tracking program. The Sparc tracking program was actually run twice, using two different pruning techniques. The HPC-tracker employed a slower pruning technique, pruning tracks which were not updated after 20 scans. This is the Sparc-tracker pruning technique that should be compared. Of course, the faster technique of pruning after 20 seconds (about 2 scans) could be added to the HPC-tracker, but this was not done, in order to let the HPC-tracker demonstrate performance increases against a larger workload. While Figure 7 shows the total time spent by the program in various intervals of the program, figure 8 shows the standard deviations of those times recorded for each run. This gives some measure of how these values fluctuated during the runs.

Conclusions

The graph in figure 5 shows that the original tracking program had near linear speedup times. It also shows that the modified version operated slightly faster, using (N+1) nodes, than the original tracking program did using N nodes. As the number of nodes increases, this difference becomes much less significant, justifying the decision to separate the supervisor/worker in order to allow efficient network communications.

Figure 7 shows that, when more than 4 nodes are used, the time lost giving up the dual role of supervisor/worker node was much less than the time saved in transmitting tracks. Clearly this provides a much faster tracker in the networked environment. Furthermore, changes in the worker process algorithms need not be duplicated in the supervisor.

Figure 7 shows that (due to the necessary serial overhead within the workers) the total time spent during the parallel portions of the program increases with the number of workers. It also shows that if one continually adds more nodes to the fixed workload, that the actual time to complete the run will increase. In the case of the tracker, optimum efficiency is achieved when

the workers finish computing association costs, just as the previous track list is sent. Since the network communication is a (nearly) fixed delay, computing resources should be added until these times coincide. Adding nodes after this point will not increase the throughput of the tracker. This is the reason for the increase in total worker time versus number of nodes.

The reason that the total run time didn't decrease monotonically until 17 nodes were used was that the times recorded were summed up over the entire run and possibly due to process switching overhead in the node on which the supervisor and detection processes ran. By queuing the detections and placing the detection client on its own node, the throughput of the system would probably increase. This would require that an additional node be dedicated for detections, which might not be the most efficient use of resources. Also with regards to maximum throughput at some loss of efficiency, one could dedicate a node to an outgoing tracking client process. The supervisor process could send the track list to the outgoing tracking client using a fast message pass, and then the tracking client would queue the track reports and transmit them over the network to the display client as quickly as possible.

Figure 7 also shows the dramatic speed increase in using different pruning techniques. The time required to perform this operation is negligible, yet it allowed the Sparc-tracker to perform faster than the HPC-tracker using 18 nodes. However, pruning tracks too quickly can cause the tracker to drop and start a new track unnecessarily, which is clearly undesirable. If the alternate pruning technique was employed on the HPC it would be able to handle a dramatic increase in workload. Furthermore, network communication would quickly become the bottleneck, such that fewer nodes would be required to track a larger, denser area.

Figure 7 (and Figure 8) shows that as the number of nodes increases, the time to complete a track cycle decreases, and the fraction of time spent in solving the 2-Dimensional assignment problem becomes larger (approaching 15%). It may be worthwhile to employ faster algorithms in this portion of the program.

The HPC-tracker operates within the improved MIST 2.0 application in an efficient, robust manner. But, there is certainly room for improvement; and some of those options have been mentioned here. Another interesting alternative is to extend the tracking program to run on a set of workstations. Since the total time spent by the worker processes is much larger than the

network communication times, and because the original tracking program did not exploit any machine-specific operations, it might be worthwhile to map the tracking program onto a a set of separate computers. Also, since the original tracking program avoided machine-specific operations, and because only one of the two application processors available in each node is currently utilized, one could (nearly) double the effective number of worker nodes by running two worker processes (using threads or otherwise) on each worker node.

References

[1] Bar-Shalom, Y. and Li, X. R., *Multi-target-Multisensor Tracking: Principles and Techniques*, YBS Publishing, 1995.

[2] Popp, R. L., Pattipatti, K. R., and Gassner, R. G., "Multitarget Tracking Parallelization for High-Performance Computing Architectures"

[3] Yeddanapudi. M., Bar-Shalom, Y. and Pattipati, K. R., "IMM estimation for Multitarget-Multisensor Air Traffic Surveillance", *IEEE Trans. AES*, 1996.

[4] Intel Paragon Programmers Manual

[5] Intel Paragon Website: http://www.hpcm.dren.net

General MIST 2.0 Block Diagram

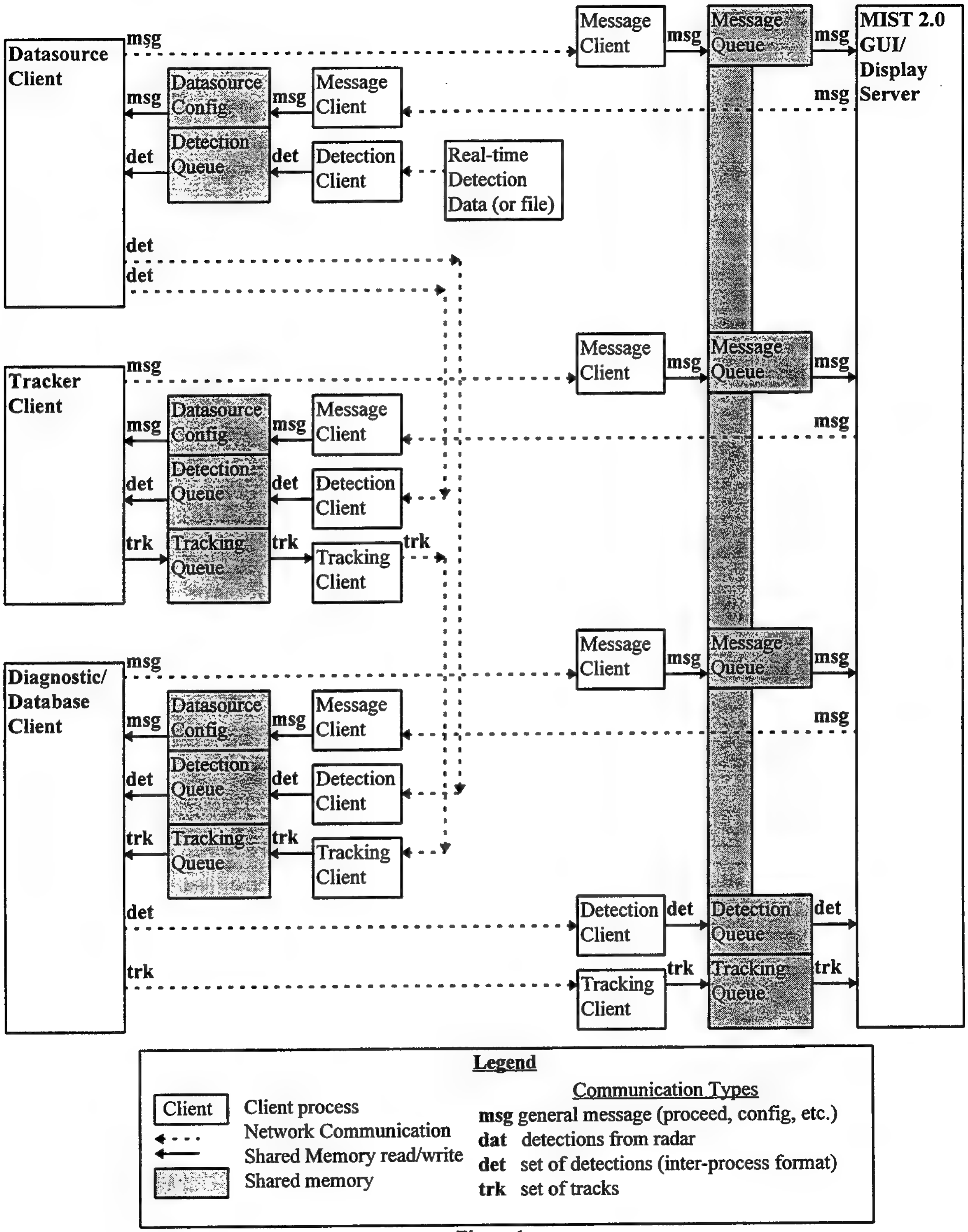
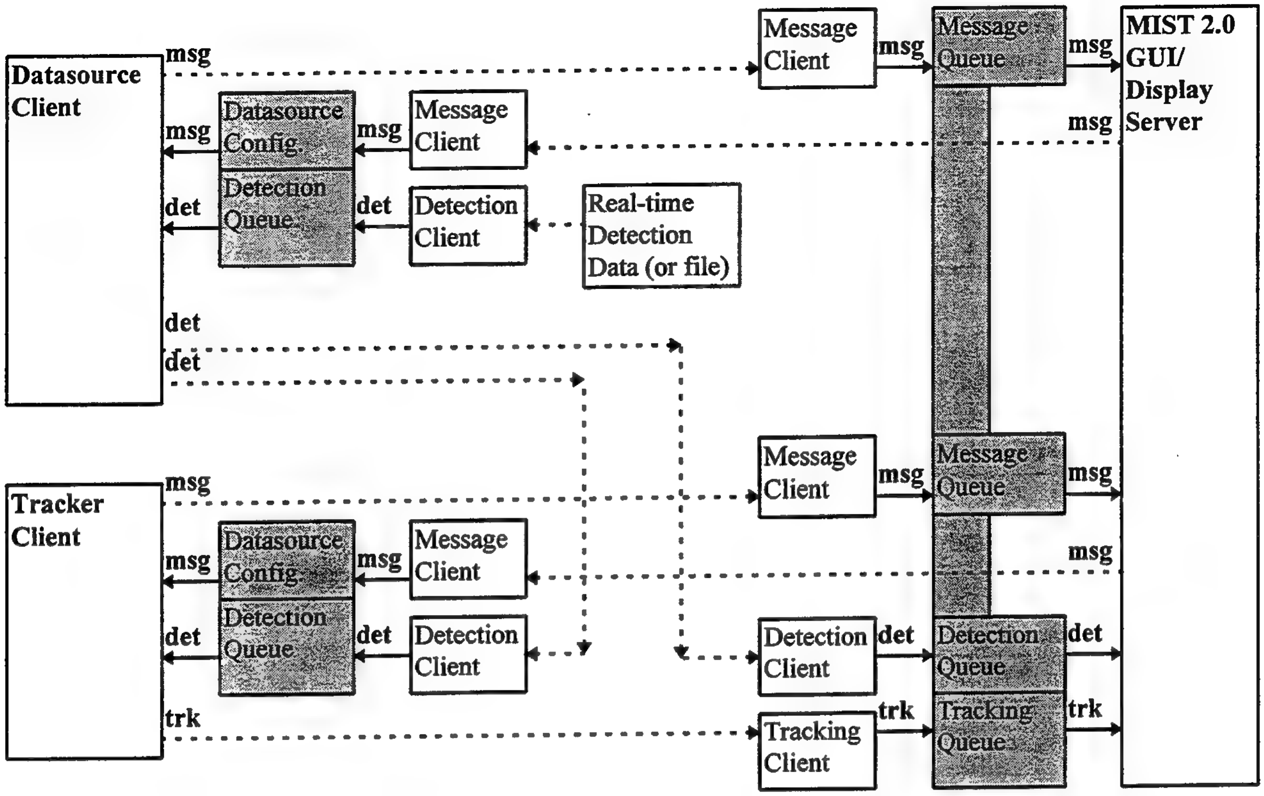


Figure 1

MIST 2.0 Components Used with Sparc-Tracker



Legend

		Communication Types	
Client	Client process	**msg**	general message (proceed, config, etc.)
← - - -	Network Communication	**dat**	detections from radar
←——	Shared Memory read/write	**det**	set of detections (inter-process format)
(shaded box)	Shared memory	**trk**	set of tracks

Figure 2

MIST 2.0 Components Used with HPC-Tracker

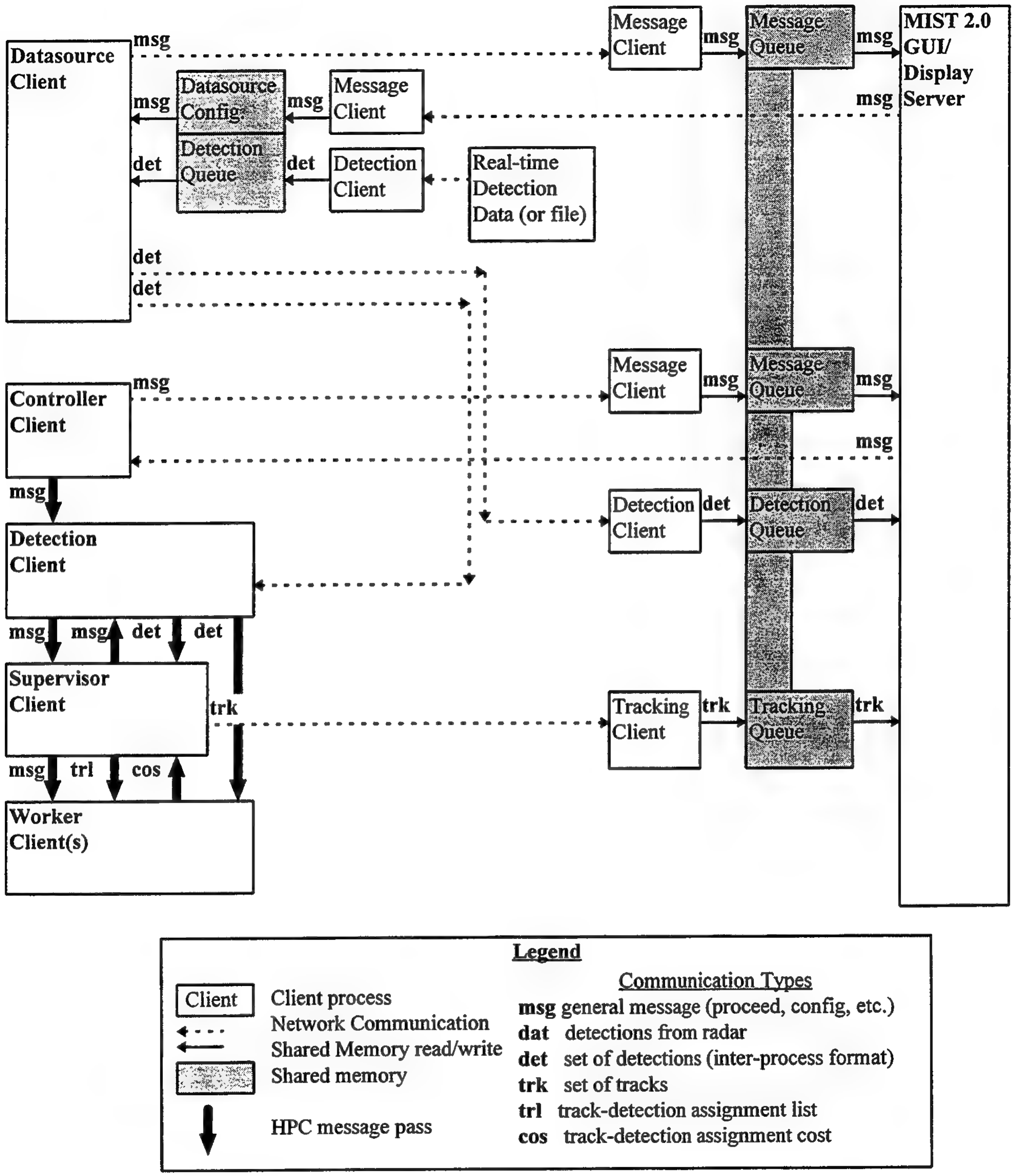
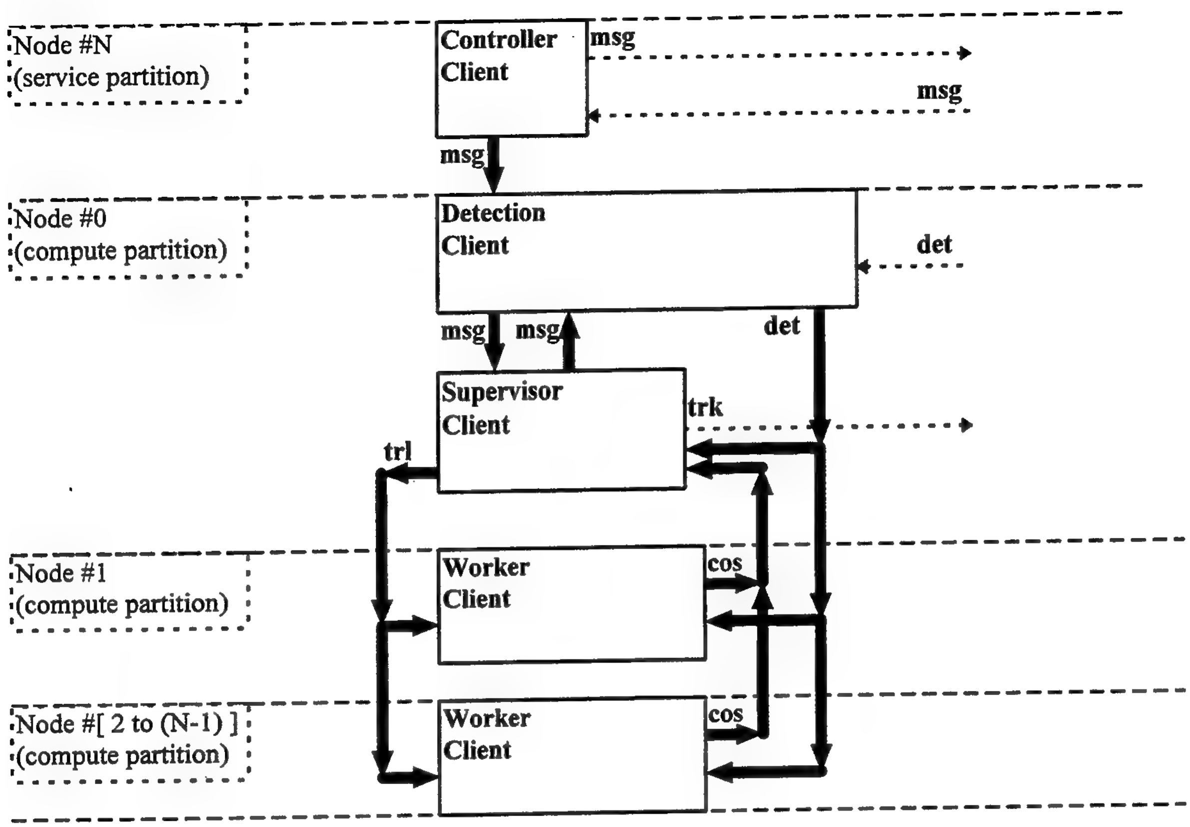


Figure 3

HPC-Tracker Process Allocation to Nodes



Legend

		Communication Types
Client	Client process	**msg** general message (proceed, config, etc.)
← - - -	Network Communication	**det** set of detections (inter-process format)
↓	HPC message pass	**trk** set of tracks
		trl track-detection assignment list
Node #	Node and Partition in which process runs	**cos** track-detection assignment cost

Figure 4

Comparison of test runs in which track formation time alone was logged:

Set #1 uses a Supervisor/Worker combined node (S/W), Set #2 does not
(note: there was no detection client process during these tests)

#nodes	Set #1 Track time with S/W	total worker-seconds with S/W	Set #2 Track time without S/W	total worker-seconds without S/W
N	T1	T1*N	T2	T2*(N-1)
51			14.76	738.00
31			21.23	636.90
26			22.18	554.50
20			27.95	531.05
15			36.40	509.60
13			40.22	482.64
12	40.62	487.44	41.16	452.76
10	45.26	452.60	50.99	458.91
9	52.56	473.04	55.17	441.36
8	56.27	450.16	64.14	448.98
7	65.29	457.03	74.19	445.14
6	72.70	436.20	86.20	431.00
5	82.03	410.15	102.69	410.76
4	104.68	418.72	131.41	394.23
3	134.71	404.13	193.73	387.46
2	196.51	393.02	376.59	376.59
1	380.23	380.23	N/A	N/A

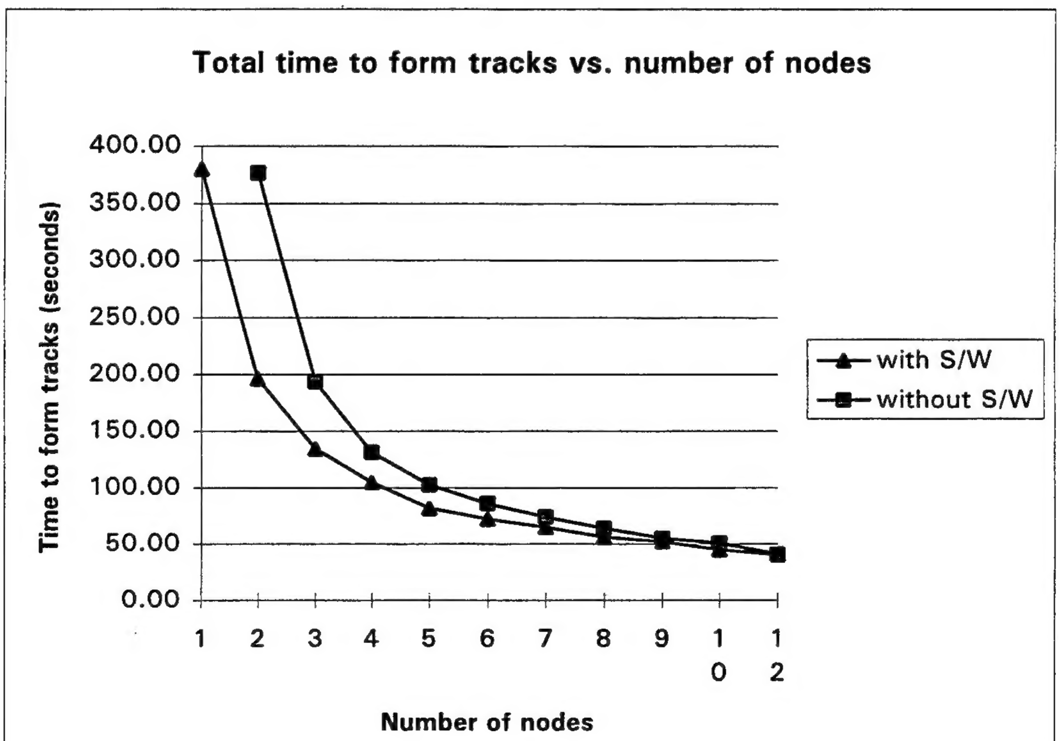


Figure 5

Timing Diagram of the HPC-Tracking program

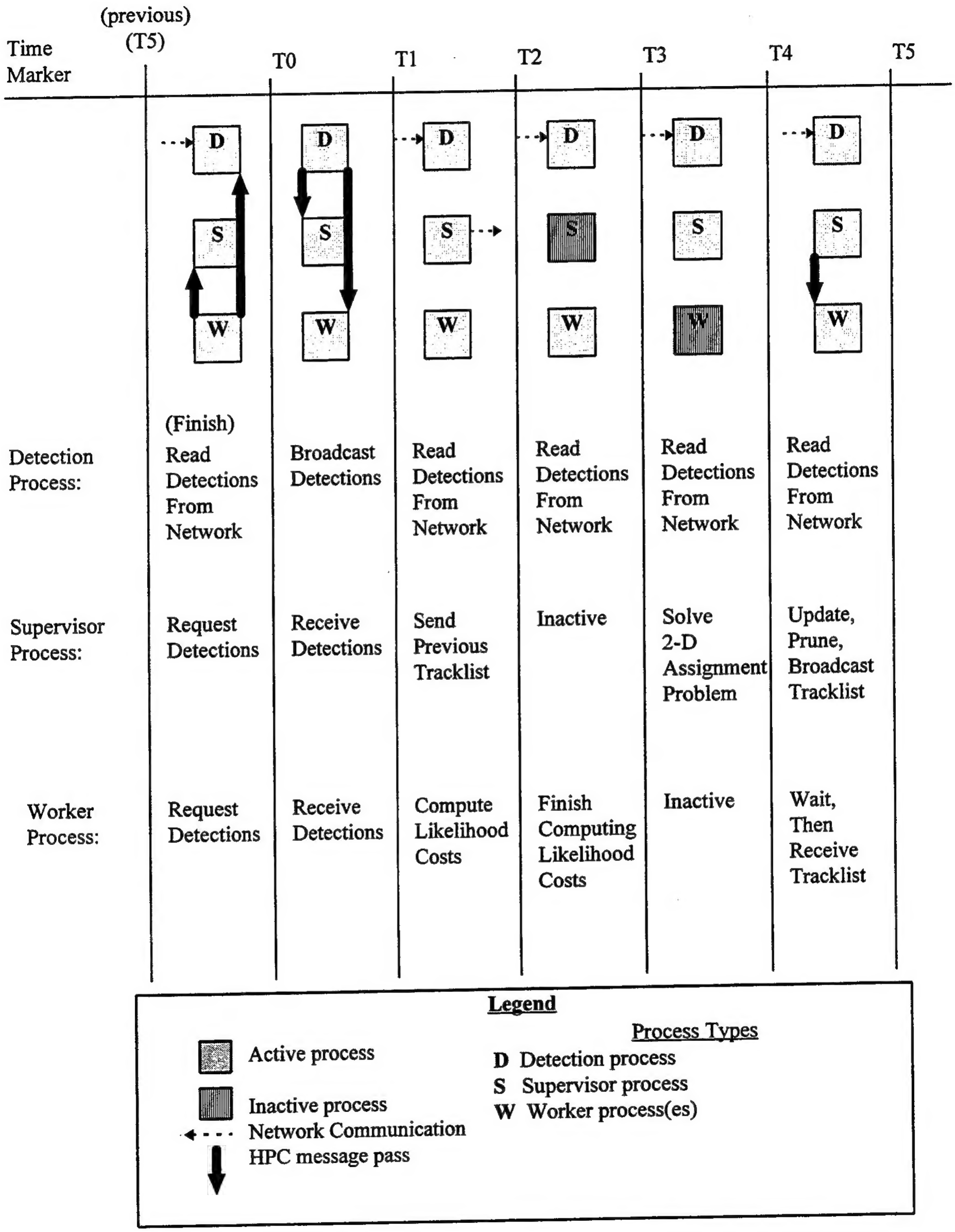


Figure 6

Comparison of HPC-tracker and Sparc-tracker performance within the MIST2 distributed application

Total #nodes N	receive detections [T0-T1]	transmit tracks [T2-T1]	finish cost calc. [T3-T2]	solve assign. [T4-T3]	update & prune trks [T5-T4]	Total time for run [T5-T0]	Total worker calc. time [T3-T1]*(N-1)
2	2.13	83.65	447.09	13.74	0.24	547	531
3	1.21	119.33	163.62	13.73	0.24	298	566
4	1.37	104.71	92.25	13.74	0.25	212	591
5	1.50	85.59	68.07	13.75	0.25	169	615
6	1.51	83.55	42.95	13.75	0.25	142	632
7	1.78	89.10	15.57	13.73	0.25	120	628
8	1.78	89.10	15.57	13.73	0.25	120	733
9	1.86	72.95	19.18	13.74	0.25	108	737
10	1.99	72.42	14.70	13.74	0.25	103	784
11	2.08	73.70	9.31	13.74	0.26	99	830
12	2.20	74.62	4.93	13.75	0.26	96	875
13	2.48	77.86	4.50	13.75	0.26	99	988
14	2.42	75.02	3.07	13.75	0.26	95	1015
15	2.51	76.76	2.07	13.75	0.27	95	1104
16	2.65	82.71	1.07	13.75	0.26	100	1257
17	2.85	85.25	0.11	13.75	0.27	102	1366
18	7.31	109.86	0.11	13.76	0.34	131	1869
19	5.13	163.49	0.11	13.73	0.38	183	2945
Additional HPC-tracker test runs, without sending tracks							
2	1.07	0.36	527.96	13.64	0.24	543	528
9	1.95	0.38	79.68	13.63	0.26	96	640
17	2.92	0.39	44.28	13.65	0.27	62	715
Sparc-20 tracker test runs							
Prune after:	[t0-t1]	[t2-t1]	[t3-t2]	[t4-t3]	[t5-t4]	total	
20 scans	0.70	2.77	0.06	0.09	502.32	506	
40 seconds	0.68	1.72	0.07	0.09	88.45	91	

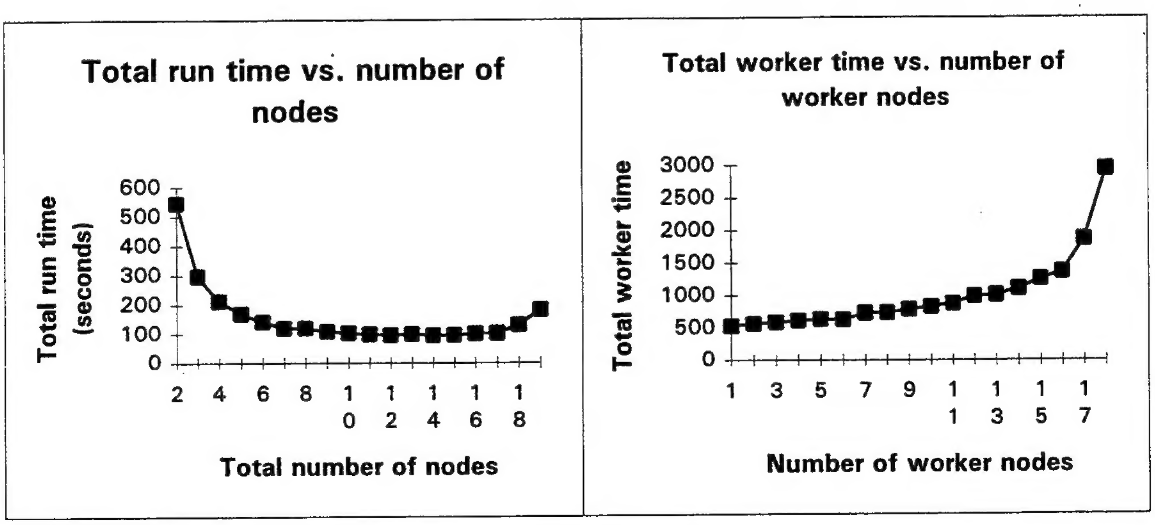


Figure 7

Comparison of the standard deviations of time intervals within each test run

Total #nodes N	receive detections [T0-T1]	transmit tracks [T2-T1]	finish cost calc. [T3-T2]	solve assign. [T4-T3]	update & prune trks [T5-T4]
2	.02	.49	6.24	.07	.0018
3	.02	1.04	2.84	.07	.0018
4	.02	.72	1.73	.07	.0019
5	.02	.62	1.29	.07	.0019
6	.01	.49	.99	.07	.0019
7	.01	.67	.49	.07	.0018
8	.01	.67	.49	.07	.0018
9	.02	.40	.52	.07	.0019
10	.01	.38	.44	.07	.0018
11	.01	.39	.30	.07	.0019
12	.01	.41	.20	.07	.0018
13	.03	.50	.19	.07	.0019
14	.01	.41	.14	.07	.0018
15	.01	.53	.13	.07	.0018
16	.01	.28	.07	.07	.0019
17	.02	.55	.00	.07	.0019
18	.34	.58	.00	.07	.0032
HPC-tracker test runs: without sending tracks					
2	.02	.00	6.64	.07	.0018
9	.02	.00	.99	.07	.0019
17	.03	.00	.53	.07	.0019
Sparc-20 tracker test runs					
Prune after:					
20 scans	.02	.01	.00	.00	6.1994
40 seconds	.01	.01	.00	.00	.6319

Figure 8